MODELING, CHARACTERIZATION, AND CONTROL OF THE ELECTRICAL-THERMAL INTERACTIONS IN ADVANCED PACKAGES

A Thesis
Presented to
The Academic Faculty

by

Wen Yueh

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2015

MODELING, CHARACTERIZATION, AND CONTROL OF THE ELECTRICAL-THERMAL INTERACTIONS IN ADVANCED PACKAGES

Approved by:

Professor Sudhakar Yalamanchili,
Committee Chair
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Saibal Mukhopadhyay,
Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Yogendra Joshi
School of Mechanical Engineering
Georgia Institute of Technology

Professor Sung Kyu Lim
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Arijit Raychowdhury
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 17 August 2015

This dissertation is gratefully dedicated to my loving parents

Peng-Hsiang Yueh and Chin-chu Yu.

ACKNOWLEDGEMENTS

I am grateful to my advisor, Professor Saibal Mukhopadhyay for guiding and advising me during the course of my Ph.D research. I am thankful for his support, encouragement, and trust in me. I also would like to thank Professor Sudkar Yalamanchili and Professor Yogendra Joshi for invaluable suggestions on my research. I also would like to thank Professor Sung Kyu Lim and Professor Arijit Raychowdhury for improving my dissertation as committee members. I would like to thank the past GREEN lab members that I worked with Minki Cho, Suhbo Chatterjee, Kwanyeob Chae, and Denny Lie for their collaboration and mentoring. I am very lucky to be working with Amit Trivedi, Khondkar Ahmed, Jae Ha Kung, Sergio Carlo, Duckhwan Kim, Jong Hwan Ko, and Faisal Amir on various exciting projects. To our group members that we did not jointly publish but greatly appreciate our time together Boris Alexandrov, Monodeep Kar, Taesik Na, and Arvind Singh for their valuable suggestions and discussions. I wish to further thank CASL lab members, William Song and Hugh Xiao for our collaboration. I wish to thank METTL lab member Zhimin Wan for our collaboration and his dedication to our project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The multi-chip integration in an advanced packaging has made the multi-physics interactions increasingly important. The objective of this research is to address the thermal coupling and the power density limitation of in-package systems through modeling, circuit emulation, and reduce thermal and power couplings. The first research objective is to construct a simulation framework to identify thermal and electrical coupling within the package. The second objective is to evaluate the sub-system's parametric failures across technologies under the influence of the coupling effects. In this analysis, memory systems with the minimal device features were studied under coupling. The framework identifies the interaction of thermal and power coupling for 2D, 2.5D, and 3D integrated coupled victims and predicts the coupling mechanism from the aggressor cores. Third, in order to refine the thermal coupling and supply characterization, a hardware emulation platform is implemented to emulate aggressors' power patterns that resembles in-package high performance cores. Along with the integrated monitoring structures, the hardware platform improves within package observability and evaluates coupling in an advanced package environments. The hardware-assisted emulation framework evaluates the package platform and supplies experimental coupling data to simulation-based systems. Lastly, a thermal-electrical evaluation on a commercial cooling integration is discussed. A full SoC is under investigation on the power, performance, and thermal interaction. The advanced package integration and the system management techniques are applied to observe on system level energy improvement through power and temperature manipulation.

CHAPTER I

INTRODUCTION

The multi-chip integration in an advanced packaging has made the multi-physics interactions increasingly important. The trend of migrating board level circuits onto a system in package (SiP) has been driven by the cost reduction needs in smaller form factor, the power reduction through fewer input-and-output (IO) communications, and a higher communication bandwidth and shorter interconnects between dies within the package. Specifically, the interposer integration and through-silicon-via (TSV) stacking drastically increases the processor-to-memory communication bandwidth, the system performance, and reduces the design footprint [40, 64]. As the interposers and TSV stacks pushing further for higher performance systems, the full system requires more rigorous design methodologies to ensure functionality. In the fields of microprocessors, embedded systems, and field-programmable gate arrays (FPGAs) the uniquely high bandwidth communication between active cores and embedded memories gain traction to improve the method of integration [33, 19].

While many in-package design parameters may be extrapolated from the traditional system board environment, some design choices are less alike to its predecessors. One of the phenomena that is unique to within-package integration is the strong physical coupling in thermal and power domains because of the short die-to-die distance within the package. Before taking advantage of the in-package designs promised, new methodology in modeling, circuit emulation, and thermal/power coupling reduction should be developed for these sub-systems to ensure functionality.

The thermal coupling and power coupling may produce undesired failure in the SiP

designs. Additional chips within the package rise the maximum power ceiling in package. Thermal coupling in package is non-trivial because the hotspot on one die may be on the heat extraction path of the surrounding dies. Throughout generation, the heat spreader surface area maintains roughly constant, the power density in advanced packaging is additive for each additional core in the package. In 3D die stacking, the within-stack hotspot become the limiting factor for system performance and power ceiling. The phenomenon is know as vertical tier-to-tier coupling. During transient operation, the unstructured workload and spatiotemporally varying power in multi-core environment makes coupling and peak power non-deterministic [34]. The variable workload in operating system scheduling, the die-to-die variation in the process parameters, and thermal-leakage interaction further complicate the predictability for adaptive design of advanced system. The sum of all iterations in the thermal coupling between die triggers unexpected thermal throttling and reduces performance. These unexpected events may be monitored and mitigated through design-time modeling and run-time monitoring circuits. When proper methodologies are in place, thermal adaptive design in a highly integrated system may improve opportunist power saving and boost operating performance beyond the traditional board level solutions.

The objective of this research is to explore methodologies to model, characterize, and control the electrical-thermal interactions in advanced packages, specifically, focusing on systems with limited cooling capacity. Three approaches have been taken on (1) the software modeling simulation, (2) the hardware system emulation, and (3) experimental system evaluation on advanced integration.

In Chapter 2 and Chapter 3 the software modeling framework is introduced. In this thesis, memory sub-systems are chosen for coupling study as they exercise the minimal design rules for a given process technology and are relatively susceptible to process variation and noise injection than the standard cells [41, 42]. The vulnerability of the memory subsystems to coupling often requires more rigor study in the margins

and failures during the design phase and has since used to study coupling electrical-thermal interaction in this work. The framework enables evaluation of advanced packaging platform during system design and evaluation. Constructing simulation framework enables understanding of the large scale 2 D, advanced 2.5 D, and 3 D stacked coupling from aggressor cores. The memory parametric failures due to the die-to-die coupling are identified through thermal and power co-simulation. Two major high performance memories are focused in this work. One is the traditional SRAM cell design, and the other is the embedded DRAM cell that has gained increasingly attention from major microprocessor manufacturers [29, 99].

In Chapter 4 a hardware system emulation framework is introduced. In order to identify core-to-core coupling in advanced packaging, transient thermal and power emulation and built-in characterization circuits are designed for advanced package evaluation. The experimental framework may be used as a transient hotspot modeling circuit for thermal conditioning and identifies the coupling victims. The multi-physics emulation framework may be used to estimate the system leakage and predict thermal behavior from power patterns and thermal coupling.

In Chapter 5 the thermal-electrical evaluation on an advanced integration is discussed. Taking a step further, a full SoC is under investigation on the power, performance, and thermal interaction. The advanced package integration and the system management techniques are applied to observe on system level energy improvement through power and temperature manipulation.

CHAPTER II

MODELING ELECTRICAL-THERMAL INTERACTION IN 3D PACKAGES – APPLICATION TO SRAM

2.1 Introduction

The chapter considers two major sources of in-package physics cross-talk for modeling. First, the heat dissipated in the cores increases the temperature within the package. Hence, the sub-system blocks closer to the hotspots also become hotter than the sub-system thats physically further away. Temperature variations modulate the device parameters thereby changing the device's characteristics – this is referred to as the thermal cross-talk. Similarly, the on-chip power delivery network (PDN) in a package is composed of planar meshes joined by the power and ground (P/G) interconnects for many dies. The in-package P/G network reduces the electrical impedance between the grids on each sub-system [60]. However, the power dissipation in the aggressor can now inject supply noise to the victim through these supply interconnects, which ultimately affects the device parametric failures. This is referred to as the supply cross-talk. Further, the thermal profile modifies the micro bumps and PDN mesh impedances. The higher temperature in the sub-system due to thermal cross-talk therefore increases the PDN impedance of the victim die, which further degrades the victim supply noise. We define this phenomenon as the tier-to-tier thermal and supply (or total) cross-talk in a in-package integration.

Through-silicon-via (TSV) based 3D stacking significantly increases the die-to-die communication bandwidth, the system performance, and reduces the design footprint [64]. The technology has received much attention in the microprocessors, embedded systems, and field-programmable gate array (FPGA) applications [33]. Many

methods have been proposed to develop TSV technologies, TSV-aware physical design tools, analyze the defect behavior of TSVs, and pre/post-bond test methods for TSV based 3D-ICs [80]. The logic cores and static random access memory (SRAM) stacking has emerged as a key application to the 3D integration. However, the supply noise and temperature variations in the 3D IC aggravate the SRAM performance and robustness in addition to manufacturing variations. In the scope of this work, we define the term "robustness" as the SRAM resiliency to parametric failures due to changes in operating condition, i.e. read failures and write failures considering variations in process, temperature, and supply noise. The additional SRAM parametric variations (e.g. threshold voltage variation) lead to failures in nanometer technology nodes, but the effect of tier-to-tier cross-talk to SRAM robustness is less understood [59, 23, 87, 100].

This chapter analyzes the performance and robustness of SRAMs within a heterogeneous 3D-stack of logic cores and SRAM arrays. The analysis on 3D-stacked SRAM considers the thermal and supply noise coupling between the heterogeneous dies. The objective is to evaluate how the logic cores (aggressors), which have higher peak power, modulate the robustness of the SRAM blocks (victim) in the 3D stack. We develop a cross-talk aware performance and robustness analysis for the 3D die-stacked SRAMs to model and analyze this phenomenon in Figure 1. Our approach evaluates the supply noise (IR-drop) and temperature of the SRAM tier considering the tier-to-tier coupling. The changes in the operating conditions are next coupled to the SRAM variability analysis considering the random dopant fluctuation (RDF) induced threshold voltage (VTH) variations. The analysis shows a strong correlation between the power dissipation in cores and the SRAM stability in a 3D IC.

The rest of the chapters are organized as follows: Section 2.2 discusses the related work and contributions of this work; Section 2.3 presents the simulation and analysis

framework; Section 2.4 presents simulation results describing various cross-talk models on SRAM performances; Section 2.5 presents simulation results under different 3D stacking orders and PDN designs; and Section 2.6 presents the chapter summary.

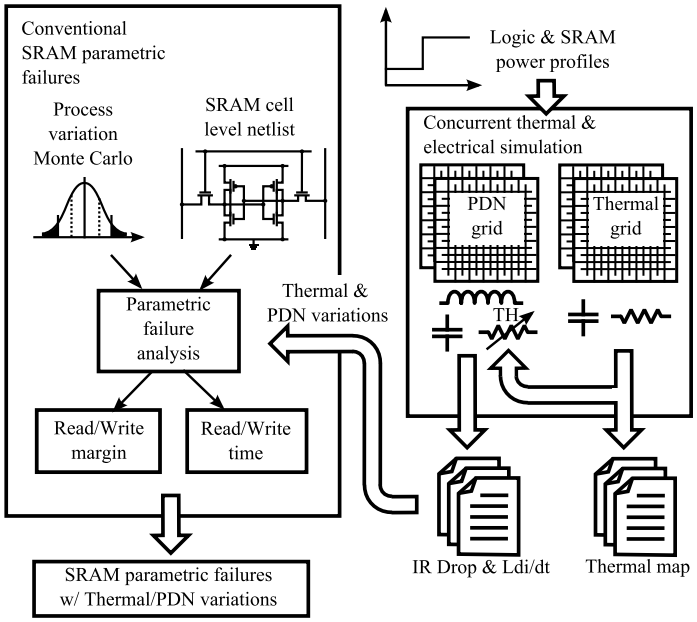


Figure 1: The simulation methodology for thermal and supply cross-talk aware SRAM analysis. The methodology co-simulates supply and thermal grids with process variation aware SRAM analysis.

2.2 Related Work

2.2.1 Electrical Observations

The 3D PDN analysis has received significant attention in the recent literature. Pak *et al.* studied PDN of a 3D-stack of graphics processing unit (GPU) cores and dynamic random-access memory (DRAM) stacks [60]. The analysis concludes that increasing the P/G TSV count improves the impedance in a limited frequency range and the effective benefit may be outside of the operating frequency. As a result, the supply noise in the DRAM could be limited by the planar PDN mesh regardless of the number of P/G TSVs. Amirali *et al.* made similar observation that, at higher TSV density, although IR drop reduces, the resistor, inductor, and capacitor (RLC) resonance quality factor increases [71]. This produces higher anti-resonances in the

high frequency range. In attempt to improve the high frequency response, Zhou *et al.* studied the use of hybrid decaps (metal-insulator-metal (MIM) decaps and traditional MOS-CAP) for 3D stacks [101]. Healy *et al.* and Tsioutsios *et al.* showed that TSV clusters improve the impedance property in a 3D die-stack [30, 82]. They identified the correlation between TSV densities and noise on the system level. Healy also studied the impacts of relative position of the high performance die within the stack, namely the core last (i.e. core closer to heat sink), core first, and core interleaved configurations. The 3D PDN research brought up concerns unique to 3D PDNs that diverge from the traditional 2D PDN restrictions.

2.2.2 Thermal Observations

Thermal modeling for 3D die-stack has also been an active area of research. The finite element methods and the computationally less complex electrical mesh based methods are often used to perform these studies [59, 82]. Models with such granularity are suitable to identify temperature distribution sensitivity versus various packaging parameters [2, 25]. These modeling techniques also quantify the die folding thermal penalty due to a higher power density [67, 68]. Much existing work uses the technique to model the thermal-electrical implications in die stacking [101, 50, 31]. A thermal framework with sufficient granularity is required to establish the proposed observations.

2.2.3 SRAM Analysis in 3D Stacks

Prior work on 3D PDN and 3D thermal models did not discuss the consequences of tier-to-tier thermal and power supply interactions on the SRAM behaviors. This work builds on the PDN and thermal modeling efforts in the previous literature to analyze the tier-to-tier cross-talk and its impacts on the robustness and performance of the 3D integrated SRAM. There are related research on 3D stack of logic cores and SRAMs [64, 33, 59, 2, 37]. However, these analyses focus on the architecture or

manufacturing aspects and not the coupled interference of power and supply cross-talk. Loi *et al.* evaluated the effect of thermal cross-talk in a 3D core-cache-memory stack, but the effect was observed from simplified lumped models [50]. Lumped tier model lacks the ability to determine planar thermal gradient which is significant to determine hotspot coupling between tiers. Hence, these analyses may not predict the 3D SRAM parametric failures due to non-uniform power distribution. A preliminary version of this work has studied the effect of supply cross-talk between core and memory on the SRAM robustness [93]. However, the combined effect of thermal and supply cross-talk were neglected. The analysis of the 3D tier-to-tier cross-talk on the SRAM stability considering the coupled impact of thermal and supply interactions is, therefore, a unique contribution of this work.

2.3 Electrical-Thermal Modeling Framework

We develop a framework to predict the SRAM robustness and performance in a 3D die-stack, considering the thermal and electrical tier-to-tier cross-talk. Since the power dissipation in cores is much higher than the SRAM's, we consider the core tiers as the aggressor and SRAM tier as the victim. Our simulation framework estimates the supply noise and temperature on the SRAM tier due to power dissipation in the processor cores, and evaluates the SRAM stability. A high level simulation flow is shown in Figure 1. The 3D PDN considers the physical design parameters such as organization of the stack, P/G TSV density, etc. The core power variations are inputs to the PDN model for both low-frequency (IR drop) and high-frequency voltage variations (Ldi/dt). Our PDN model uses a distributed RLC grid (details described later) and simulates the circuit using HSPICE in Figure 2(a). Similarly our thermal framework transforms thermal components into an equivalent distributed RC model and uses HSPICE as the backend simulator in Figure 2(c). The resistances of wires and TSVs in the PDN network are modeled as voltage controlled resistances

to couple the thermal and PDN simulations. Hence, in the simulation framework, the thermal components co-simulates with the PDN components and produce the thermal feedback directly to the planar PDN meshes and TSVs during the simulation in Figure 2(b). The recorded voltage noise affects the SRAM robustness under threshold voltage variations. We consider read margin, write margin, read time, and write time as the metrics for SRAM robustness [56]. The traditional Monte-Carlo simulation is used to generate variations of the SRAM robustness metrics considering random threshold voltages variations for various supply noise conditions and temperature conditions. The spatial coordinate on each node maps the supply voltages and operating temperatures to the locations in the aggressor cores and the victim SRAM blocks.

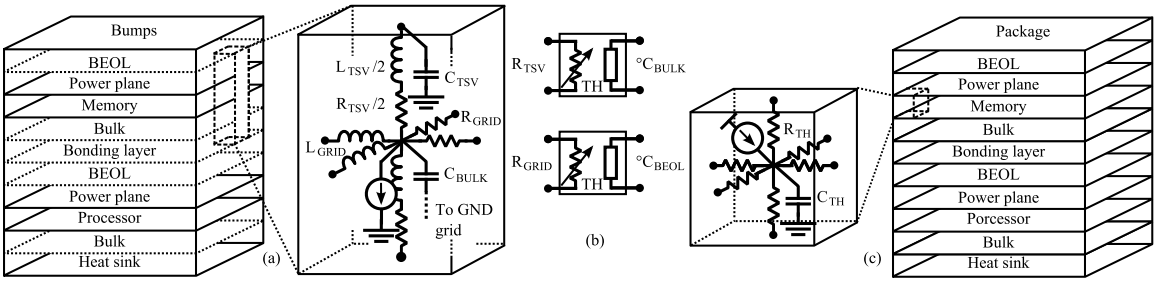


Figure 2: The thermal and supply grid model considers feedback from the thermal effects. The PDN supply unit cell and the corresponding layers it spans are shown on (a). The thermal grid unit cell and the corresponding layer it represents are shown on (c). The resistances of the PDN grid are coupled to the local temperature (b). The tier PDN resistances are coupled to the thermal BEOL layer and the TSVs are coupled with the thermal bulk layer.

2.3.1 Thermal Modeling with 3D Distributed RC Grid

The thermal framework in this study uses distributed RC grid where R represents the thermal resistance and C represents the specific heat. The grid for the planar tier includes the impact of heat sink, silicon, and insulator layers. The stack is face-to-back bonded with processor-memory tiers, as shown in Figure 2(c). The 3D stack consists of the thermal package (i.e. heat sink, spreader, and thermal interface material), the

bulk silicon, the active silicon (processor), the back end of the line (BEOL) including metal connections (for the processor layer), the die-to-die interface material, the bulk silicon, the active silicon (memory), the BEOL layer (for the memory layer), and the electrical substrate/package (i.e. die-to-package interface). We simulated a 12 mm $\times$ 12 mm chip having the same grid density as the PDN grid (48 $\times$ 48 nodes). The total thickness of the die is 350 μm. The bonding material, heat sink, and package conductivities are from the values reported in [9]. The referenced parameters including the layer thicknesses are tabulated in Table 1. The BEOL uses oxide to metal ratio of 1:3 for its thermal resistance. The thermal resistivity of the die-to-die interface layer (core and SRAM) was modified to consider the effect of the heat flow through the TSVs. The thermal power profile ratio between the core tier and the SRAM tier maintains the 10% rule; the same as the electrical system. It is chosen to be a reasonable L2 cache power figure for a microprocessor [22].

Table 1: Parameters for Thermal Simulation

Parameters	Thickness (m)	R (W/m $\bullet$ K)	C (J/m^3 $\bullet$ K)
PKG	1 m	20	35.5 K
BULK	100 μ	100	1.75 M
DEV	20 μ	100	1.75 M
BEOL	50 μ	40	4.00 M
BOND	10 μ	100	4.00 M
SINK	1 m	400	35.5 K

2.3.2 Power Delivery Network Modeling with 3D RLC Power Grid

For a 2D design, a planar power grid delivers the power directly to the SRAM cells. The PDN in a 3D die-stack is composed of multiple planar power meshes connected through vertical P/G TSVs in Figure 2(a). One 2D grid provides power to the SRAM array and the second 2D grid provides power to the cores. The currents for the cores and SRAMs are modeled with distributed current sources. The SRAM power density is the same as the thermal model and consumes 10% of processor power. The 2D

grid design is motivated from the work of Gupta *et al* [25]. The RLC network uses an equivalent distributed power mesh derived from the lumped impedance model of a Pentium 4 processor. The grid has 48 × 48 grid nodes for VDD and corresponding 48 × 48 grid nodes for ground. The die dimension is 12 mm × 12 mm and forms unit cell dimension of 250 μm × 250 μm. Note this model does not suggest the grid metal-to-metal mesh pitch is 250 μm, rather we derived an equivalent impedance model for unit cell size of 250 μm × 250 μm. The equivalent grid resistance and inductance resembles the 50 μm grid pitch mesh model used by Pak *et al.* [60]. The off-chip impedances are modeled with RLC ladders as well to capture the low frequency noise in Figure 3. The first segment of the ladder models the board level lump impedance and the second segment of the ladder modes the package impedance. The package ladder is evenly distributed to points on the on-die grid with partially-lumped controlled collapse chip connection (C4) bump impedances. A list of the grid and TSV parameters are shown in Table 2.

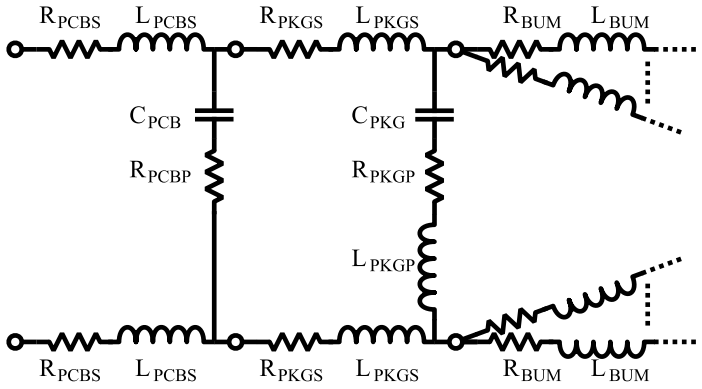


Figure 3: The off-chip impedances used in the supply modeling framework. The model contains RLC ladders for board, package, and bump distributions.

The 3D grid network uses similar planar distributed grid structure but has vertical P/G TSV connections for each stratum. Half of the equivalent P/G TSV structure is shown in the unit cell in Figure 2(a). The physical geometry of the P/G TSVs is 100 μm in length, liner thickness of 100 nm, and 10 μm in diameter. The minimum pitch is assumed to be 100 μm, and the corresponding TSV density would be 100

TSVs/mm^2. In the grid design, this is equivalent to roughly maximum six TSVs per unit grid. Due to the TSV count being higher than the grid nodes, the TSV impedance is lumped to an equivalent model during the analysis. The corresponding TSV resistances and capacitances are extracted from 3D TCAD simulation. The TCAD simulation accurately accounts for the MOS capacitance of the P/G TSVs. The in-depth P/G TSV capacitance modeling may be found in our earlier work [93, 81]. The TSV inductance is referenced from the work of Katti *et al* [36]. The 3D as well as the 2D equivalent impedances used throughout the chapter are given in Table 3.

2.3.3 Temperature Dependent Grid Model

The temperature variation changes the PDN electrical condition as the thermal grid and supply grid receive the same power profile. Using the voltage controlled resistances for the 2D grids and TSVs, the supply impedance receives the thermal condition update. Increasing PDN temperature increases PDN resistivity due to electron mean free path collisions. The on-chip PDN resistances are temperature dependent and the corresponding function uses a linear approximation,

$$R = R_0 \left[1 + \alpha \left(T - T_0\right)\right] \tag{1}$$

. The PDN temperature coefficient α is chosen to match copper's coefficient 3.9 $\times$ $10^{-3}/\circ$C. The ambient temperature T_0 is chosen to be 25°C and the corresponding R_0s

Table 2: Parameters for Grid Mesh and TSVs

Components	Geometries
Chip Dimensions	12 mm x 12 mm
P/G Grid Nodes	48 x 48 nodes
Die Thickness	350 μm
TSV Min. Pitch	>100 μm
TSV Diameter	10 μm
TSV Liner Thickness	100 nm
TSV Liner Thickness	100 nm

are included in Table 3. The off chip components maintain ambient temperature, and the on chip components heat up according to the chip power. The TSV temperatures are sampled from the thermal bulk layer in the thermal grid, and the 2D planar grid temperatures are sampled from the BEOL layers in the thermal grid.

2.3.4 Modeling SRAM's Electrical Characteristics

In SRAM parametric analysis, the predictive 32 nm technology models are selected for the simulation [9]. We evaluate the traditional 6-T cell parameters. Table 4 shows the design parameters for the cell. Figure 4 describes the basic read and write operations of the SRAM cell. During writing, one of the bitline is reduced to '0' and the other one is held high at VDD. Once the word-line is turned on, the node storing '1' is discharged through the access transistor (i.e. the NMOS pass transistor in Figure 4(a)) while the node storing '0' is charged. Once the internal node voltages cross each other, the positive feedback of the cross-coupled inverters amplifies the voltage difference to the rail-to-rail swing and the cell nodes reaches VDD and '0' [55]. A write failure occurs if the cell nodes cannot change their state within the word-line 'on' period as shown in Figure 4(a) [54]. The read operation is shown in Fig, 4b. During read operation, when the word line is raised high, the bit-line connected to the cell node storing '0' discharges through the read path consisting of the access transistor and the pull-down NMOS transistor. The bit-line connected to the node storing '1' remains high creating a bit-differential between the two bit-lines. Once the

Table 3: Parameters for PDN Simulation

Parameters	R0 (Ω)	L (H)	C (F)
PCB	94 μ (s)	21 p	240 μ
	166.6 μ (p)		
PKG	1000 μ (s)	120 p	26 μ
	541.5 μ (p)		
BUMP	40 m	72 p	
GRID	28.1 m	3.1 f	93.8 p
TSV	7.735 μ	5.710 p	313.2 f

sufficient bit-differential is developed, the sense-amplifier amplifies the difference to rail-to-rail swing. If the discharge process is slow, then the developed bit-differential is less, leading to incorrect sensing by the sense amplifier. This is known as the access failure. During the reading process the node storing '0' rises to an intermediate voltage level, defined as the read disturb voltage (Vread). If this read disturb voltage is higher than the trip-point of the inverter associated with the node storing'1', the cell content flips as illustrated in Figure 4(b). This is known as the read disturb failure. The following metrics are used to the parametric stabilities under inspection:

Table 4: SRAM Parameters

Transistors	V_{TH0} (v)	$\sigma^2(V_{TH0})$	Width (nm)	Length (nm)
PULLUP	-0.58	0.05	64	32
ACCESS	0.63	0.035	130	32
PULLDN	0.63	0.04	100	32

The bit-line write margin: The write margin is measured as the maximum voltage at the low-bit line (BL in Figure 4(a)) that results in successful write operations (the weak-write test). The write margin is associated with write failure. The write time: The write time is measured as the difference between the time-instant the wordline is raised high and the time instant the cell nodes cross each other. The write time is related to the write failure. The static read margin: The read margin is estimated as the difference between the voltage rise in the node storing '0' (V_{read}) (node $\bar{Q}$ in 4(b)) and the trip point of the inverter associated with the node storing '1' (V_{trip}) (node Q in Figure 4(b)). The read margin is used to predict the read disturb failure. The read time: The read time is measured as the time required developing 100 mV voltage differential across the bit-lines (BL and ($\bar{BL}$) in 4(b)). The read time is related to the access failure. The Monte-Carlo analysis on SRAM process variations can be found in the literature [87, 20]. The voltage and temperature variations assume all cells within a grid node in the PDN and thermal models have the same supply noise and temperature. Figure 5 includes the sweep for a cell's read margin, write

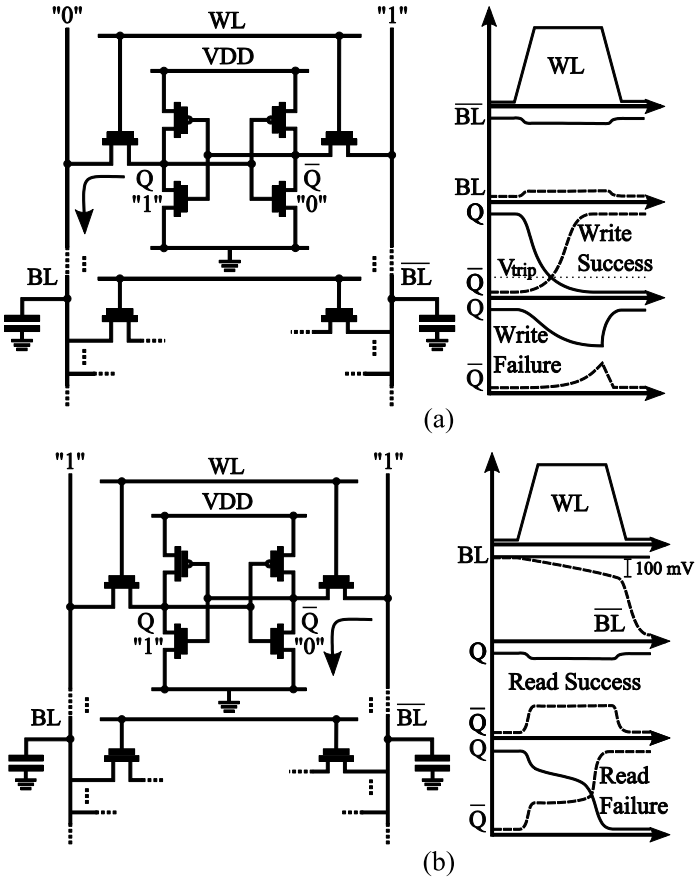


Figure 4: Simplified SRAM read and write operations. (a) shows the write operation discharging the internal node Q. A successful write flips the cell content, and a write failure retains the stored value. (b) shows the read operation discharging the ($\overline{\text{BL}}$). A successful read discharge the bit-line at the node storing zero, and a read failure flip the cell content.

margin, read time, and write time. The plot indicates the mean read/write margins are more sensitive to voltage variations than the temperature variations confirming with prior work observations [77, 94]. The sample mean of the read or write time, on the other hand, is modified by both voltage and temperature variations. We further note that the write time is observed to less than the read time in Figure 5(a) and Figure 5(b). This is because: during write process first we need to discharge the node storing '1' to a low level such that the voltages at the internal cell nodes become equal (in igure 4(a). Therefore, during the writing process only the capacitances within the internal cell are discharged (and charged). But during reading, the entire bit-line capacitance needs to be discharged to create the required bit-differential in Figure 4(b). Although the bit-line capacitance (summation of the metal capacitances and junction capacitances of all access transistors in a given column) is much larger than the capacitances of the internal cell nodes, the same access transistor regulates the discharging current in both cases. This explains that the write time is much smaller than the read time under the nominal condition. We next observe that the write time is less sensitive to thermal and supply variations compared to the read time. This is because: write time is determined by the contention between pull-up and access transistor; higher temperature or reduced voltage weakens both pull-up and access devices and weaker pull-up partially mitigates the effect of weaker access transistors. The above observations confirm compelling reason to perform a co-analysis considering the electrical and thermal properties of SRAM parametric variations within the 3D IC.

2.4 Simulation and Discussion on Coupling Analysis

The performance and robustness of SRAM are subject to 3D tier-to-tier couplings (thermal and supply). We study the robustness and performance of the 3D integrated SRAMs under the power dissipation cross-talk from cores with the simulation

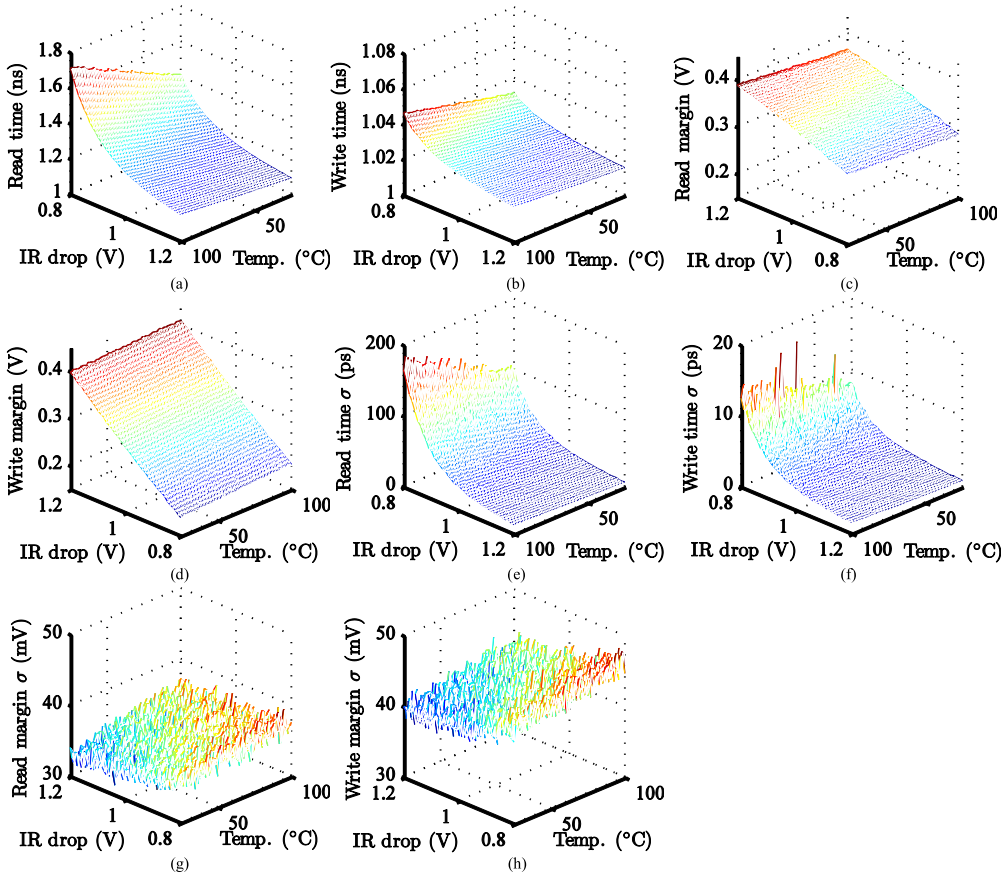


Figure 5: Sweeps on the thermal and supply conditions are performed on the SRAM cell for read time (a)(e), write time (b)(f), read margin (c)(g), and write margin (d)(h). The figures (a)–(d) are the mean shifts due to temperature and PDN conditions, and (e)–(h) are the standard deviation change due to these effects. The read time is very dependent on the thermal and supply conditions while the write time is relatively independent. The read margin has a negative dependency (at 25 °C is 340 mV and 100 °C is 320 mV keeping supply at 1.0 V) to the temperature and the write margin has a positive dependency.

framework developed in Section 2.3. The analysis focuses on a 3D stack with shared PDN design between cores and SRAMs. The SRAM tier in the stack is further from the heat sink and closer to the C4 bumps. We will refer to this configuration as the shared PDN w/ core near HS. A cartoon PDN grid representation and the corresponding stack configuration are illustrated in Figure 6(a) for visual references. The 3D grid model sustains 10% IR drop for 72 watts uniform power on the core tier, and sustains 11% IR drop for additional 7 watts on the SRAM tier. The shared PDN gird contains 4608 power and ground TSVs and 288 uniformly distributed C4 bumps shared between the SRAM tier and core tier accordingly. A core power profile and the corresponding transient variations in the temperature and supply voltage of the SRAM tier are shown in Figure 6(b).

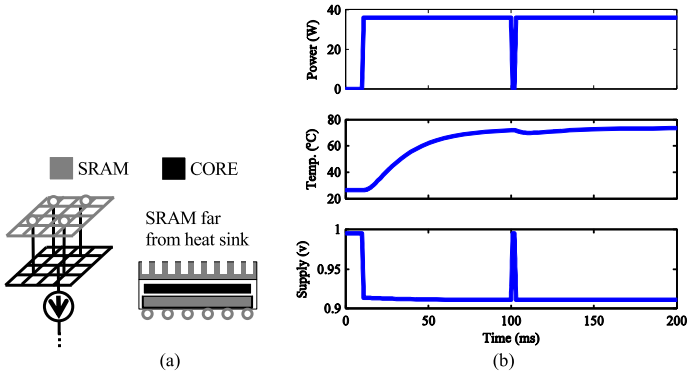


Figure 6: Representative PDN structure, stacking order, and simulation power profile: (a) shows a cartoon representation of the shared PDN connection and the package configuration (b) the power profile on the core tier affects the temperature and supply simultaneously on the SRAM tier. At 100 ms, the core power is switched on and off to simulate the Ldi/dt noise for the transient analysis.

The analysis on SRAM parametric failures in this section are discussed in the following order. First, Section 2.4.1 discusses the effect of proximity to the power source on the SRAM parametric failures. Next, Section 2.4.2 discusses the effect of maximum core power on the SRAM robustness. Then, Section 2.4.3 discusses the relation between the core powers, hotspot sizes, and SRAM robustness metrics. Finally, Section 2.4.4 discusses the additional effect of the Ldi/dt noise to the SRAM

robustness metrics. For figures in legends and labels, we use the following acronyms: RDTM – the read time, WRTM – write time, RDMGN – read margin, and WRMGN – write margin. We report the means and the standard deviations for all four metrics.

2.4.1 Effect of Power Source Proximity

On the framework described above, we consider an 18 mm^2 hotspot source dissipating 32 watts of power at the center on the core tier in Figure 7(a),(b). The steady state temperature and supply noise are measured at the SRAM tier. This studies the SRAM reliability when it is near or far from a hotspot. The IR drop on each grid point forms a supply gradient along the planar radius in Figure 7(a). Similarly, the temperature gradients are also functions of the distance, Figure 7(b). The SRAM variations under coupling are presented in Figure 8. The analysis considers SRAM cell location at center (distance at 0 mm to the hotspot) and edge (i.e. at a distance of 6 mm). First, the combined effect of the supply and thermal cross-talk significantly increases the nominal value (13%) and spread (200%) in the read time of the SRAM at the center compared to the respective process variation baselines. It demonstrates the need for a coupled analysis, not just supply or thermal coupling alone, because by themselves the mean shifts are between 5-6% and spreads increase by 94–96%. Second, the SRAM read time at the center has higher nominal read time and larger variation compared to the cell at the edge of the chip, showing a 11% different in mean and 160% spread between these two extreme locations. The proximity of the coupling hotspot is responsible for this variation, which the lumped RLC model may not identify. Also, the proximity does not affect SRAM supply coupling and thermal coupling equally. For example, Figure 8(a) and 8(b) show that the edge read time under supply cross-talk increases by 3% and the spread by 43%, while the corresponding changes are less than 1% and 8% for the thermal cross-talk alone. This suggests electrical coupling is widespread but distributed on grid, while the thermal coupling is localized but intense

near the hotspot. The write-time performance variation in Figure 8(a) and Figure 8(b) exhibits similar trend. The overall variation is, however, much less compared to the read time. The baseline process variation induces less than the read time variation and additional thermal and supply cross-talk have relative less impact. The thermal and supply variations on cell margins are illustrated in Figure 8(c) and 8(d). The combined effect of the supply and thermal cross-talk reduces the center read margin by 10% and write margin by 13% for SRAM cells near the hotspot. Note that the write margin is comparably less sensitive to the thermal cross-talk than the read margin while relationship reverses under the supply cross-talk. Note Figure 8(c) shows that the write margin of the cells vertically close to the core power is marginally better than the value considering only the supply cross-talk. The effect is due to better write-margin of the cell at higher temperature as shown in Figure 4(a). The write margin improvement may be attributed to the transistors' threshold voltage reduction at higher temperature as lower threshold voltage of the transistor improves write margin [66, 54, 53]. Also, at higher temperature the NMOS leakage current increases and assists the writing process. When thermal cross-talk is considered along with the supply cross-talk, the cells near the hotspot operate at a higher temperature. Hence, the write margins of those cells are observed to be higher than the scenario when only supply cross-talk (i.e. the SRAM at the room temperature) is considered.

2.4.2 Effect of Core Power Variation

The hotspot power determines the conditions of the coupled SRAM. We consider the thermal and supply conditions of a SRAM cell vertically aligned to a hotspot source and vary the 18 mm^2 hotspot power obtaining the IR drops and temperature profiles as functions of the core power. The simulation for 36 watts power profile and the 3.6 watts power profile are reported in Figure 9 to illustrate the hotspot power influences. Higher power dissipation in the cores creates a larger voltage drop in the

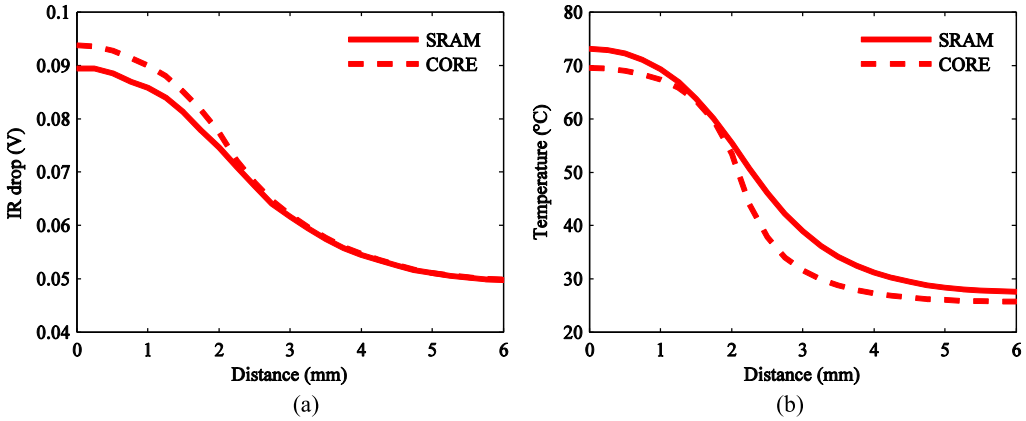


Figure 7: Representative IR drop and temperature variation along the planar distance to the hotspot source. A non-uniformed 36 watts hotspot power of 18 mm^2 is applied to the center of the shared PDN w/ core near HS design: (a) the shared PDN couples the SRAM IR drop to the core IR drop; (b) the core and SRAM couples thermally due to the heat gradient across the 3D stack.

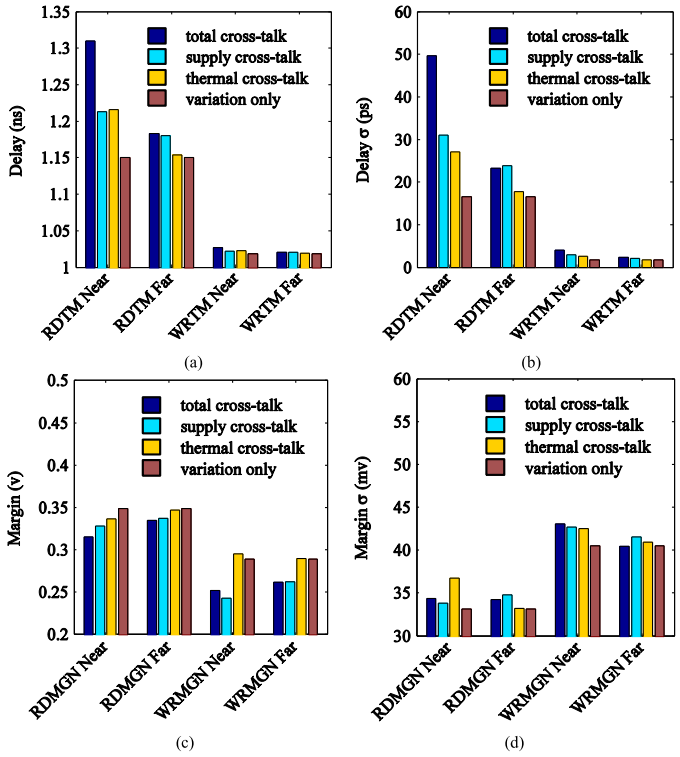


Figure 8: Comparison of the proximity cross-talk effects including the thermal cross-talk only, electrical cross-talk only, and the combined cross-talk: (a) the delay metrics of the read time and write time with the proximity distance near (vertically aligned with the hotspot) and far (at the chip edge), (b) the standard deviation of the read time and write time, (c) the mean read margin and write margin, and (d) the standard deviation of the read margin and write margin.

SRAM tier. Similarly higher power dissipation also heats up and modulates device operating point. The mean read time difference between 36 W core power versus 3.6 W power are 5% considering supply cross-talk, 5% considering thermal cross-talk, and 13% considering combined cross-talk. The standard deviation in the worst case read time increases by 188% considering the combined variations. The write time analysis resembles the proximity experiment and is weakly affected by the thermal and supply cross-talk. The higher voltage drop and temperature in the SRAM tier due to the higher core power translate to reduced margins in Figure 9(c),(d) is similar to the observation made in Figure 8(c),(d). We therefore observe a unique correlation between power in the core tier, and the SRAM parametric failures. Note the power dissipation in the logic cores is time varying and workload dependent. These factors modulate SRAM parametric variation accordingly.

2.4.3 Hotspot Limitation on Allowable Power

In this section, we study how the changes in the SRAM performance and robustness parameters impose limits on the power budget on the core tier. Figure 10(b) shows the maximum allowable IR drop and temperature contour that ensures the performance and robustness parameters (read time, read margin, and write margin) remain within 15% of their nominal value. These curves may be obtained by fixing the 15% nominal value on the z-axis in Figure 4 (the write time analysis is not included in Figure 10(b) for simplicity). We experiment with different hotspot sizes on the same coordinate: 144 mm^2 (full chip), 72 mm^2, 48 mm^2, and 18 mm^2. For each hotspot size, the thermal condition and the supply drop are subject to the hotspot power sweep. The intersections between Figure 10(a) and Figure 10(b) determine the maximum allowable hotspot powers given their respective sizes.

The maximum power limit on the core tier depends on SRAM thermal and supply cross-talk analysis. The allowable maximum power reduces as the size of the hot spot

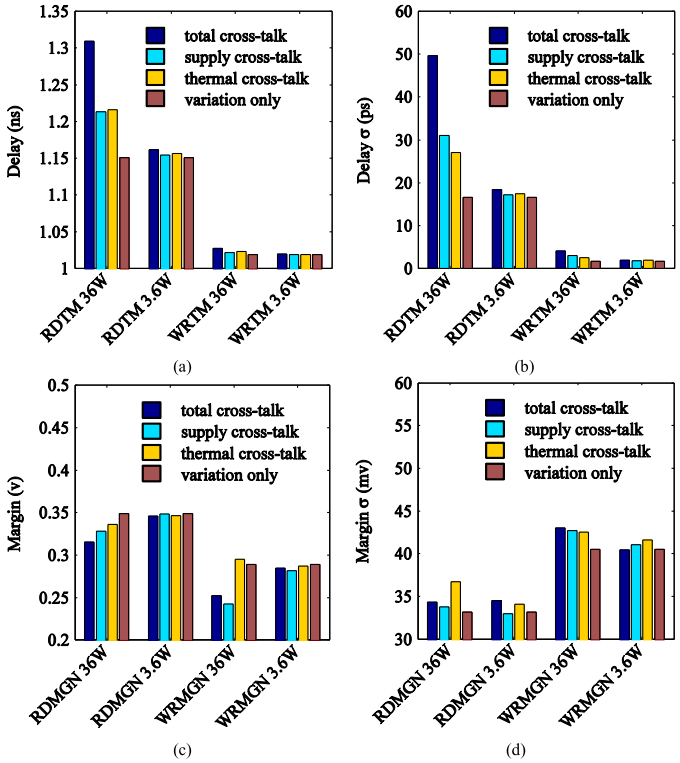


Figure 9: Comparison of the cross-talk effects due to the power variation in the source including the thermal cross-talk only, electrical cross-talk only, and the combined effect: (a) the delay metrics of the read time and write time comparing the 36 watts and 3.6 watts of power on the process tier, (b) the standard deviation of the read time and write time, (c) the mean read margin and write margin, and (d) the standard deviation of the read margin and write margin.

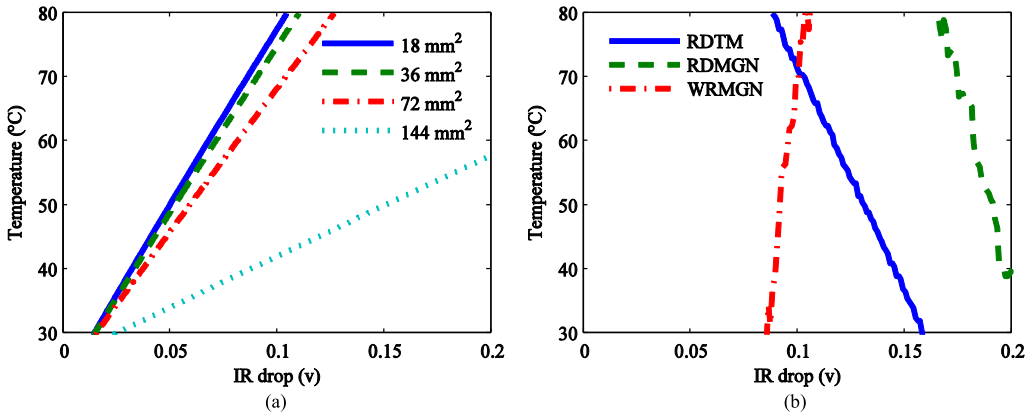


Figure 10: Different power densities produce different IR drop vs. temperature correlation in (a), and similarly 15 % nominal read time, read margin, and write time contours bound the SRAM operating conditions in (b). The intersection between each power density in (a) and the contour lines in (b) is the maximum allowable temperature and IR drop for the particular power density.

reduces (i.e. power density increases). Given a tolerance budget of 15 % off the nominal, the change in the read time imposes a stringent limit when power density is high (i.e. for the smaller size of hotspots) while write margin imposes a more stringent limit when power is uniformly distributed in Figure 11. This is because at higher power density, with increasing core power, temperature at the SRAM tier increases at a faster rate than the IR drop. Since higher temperature has a stronger impact on the read time than the write margin, the read time limits the core power in the stack. On the other hand, at uniform core power (i.e. lower power density) the IR drop is more critical than the temperature change. As the write margin has a higher sensitivity to the IR drop, the write margin limits the core power. The above analysis illustrates the thermal cross-talk, supply cross-talk, core tier power density, and SRAM parametric characteristics are important factors to determine maximum core power in the stack as well as to constrain SRAM parametric variations.

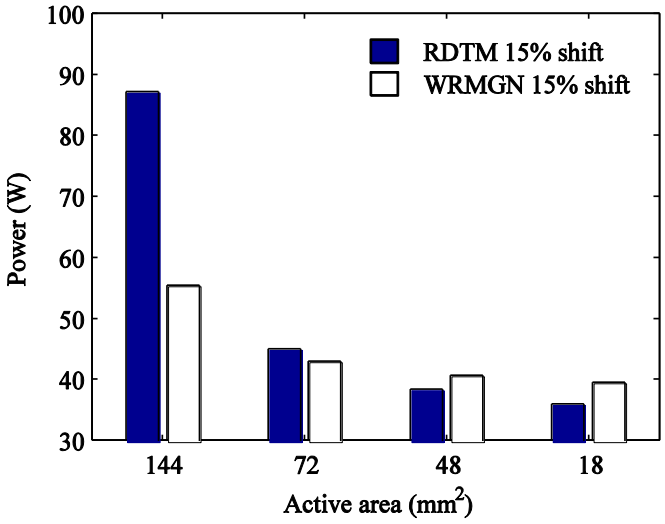


Figure 11: The allowable hotspot power derived from the allowable coupling temperature and IR drop conditions. Our SRAM design bound higher power density core tier hotspots (48 mm^2 and 18 mm^2) by its read time and distributed power by its write margin (144 mm^2 and 72 mm^2).

2.4.4 Effect of Ldi/dt Noise through Supply

High frequency Ldi/dt noise across shared 3D PDN degrades SRAM parametric failures as well. We capture the maximum droop through reducing the hotspot power on the core tier briefly by 400 ns and reactivating it immediately in Figure 12(a). The 36 watts supply pulse delivers the worst case IR drop plus the Ldi/dt noise, while being brief enough without disturbing the steady state temperature at 73 °C. The first droop near 600 ns in Figure 12(b) is a representative Ldi/dt noise additional to the IR drop. The PDN simulation with thermal coupling produce 3% deeper first droop than the PDN simulation at nominal temperature. This phenomenon may only be captured through thermal modulated PDN simulation, which independent thermal and supply simulations may not observe. The methodology is also more accurate than a temperature sweep because each PDN grid resistance receives thermal update depending on the hotspot power and proximity to the source. Because the droop period last for few nanoseconds, this voltage condition maps the worst-case SRAM parametric failures similar to the method in Section 2.4.2. The SRAM parametric failure analysis considers the coupling proximity, IR drop, worst case Ldi/dt noise, and temperature in this section. First, the SRAM read time under high frequency coupling increases the statistical spread. The supply droop increases the mean read time by 5% and spread by 87% compared to analysis with IR drop only in Figure 13(a),(b). The read time further increases to 16% and 273% for the mean and spread when the thermal effect is considered (affecting both the PDN first droop and the SRAM transistors). Second, the supply coupling noise between PDNs travels far from the source. This is because of the reduced resistance in the planar PDN mesh. The combined effect produces noticeable read time delay (6%) at the edge of the chip. The magnitude of the degradation also depends on whether the parameter is a stronger function of the supply variation. The write margin, which has a stronger dependency to the supply cross-talk in this SRAM cell, shifts its mean by 17–21% in Figure 13(c). The read

margin, on the other hand, shift by only by 7–13%.

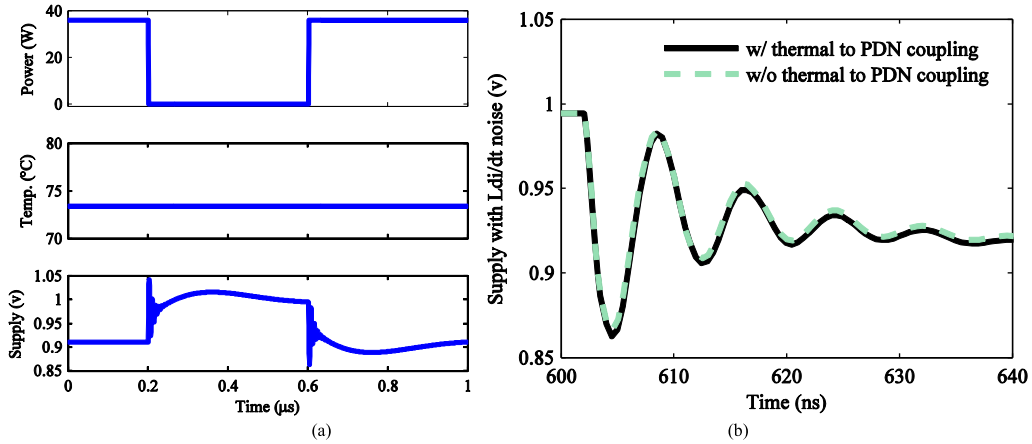


Figure 12: The supply with Ldi/dt noise illustrated the simulation of (a) the source is disabled brifly and turned on for maximum Ldi/dt droop while maintaining approximately the same temperature; (b) a zoomed view at 500 ns showing the first droop and the comparison to the simulation without the thermal to PDN coupling.

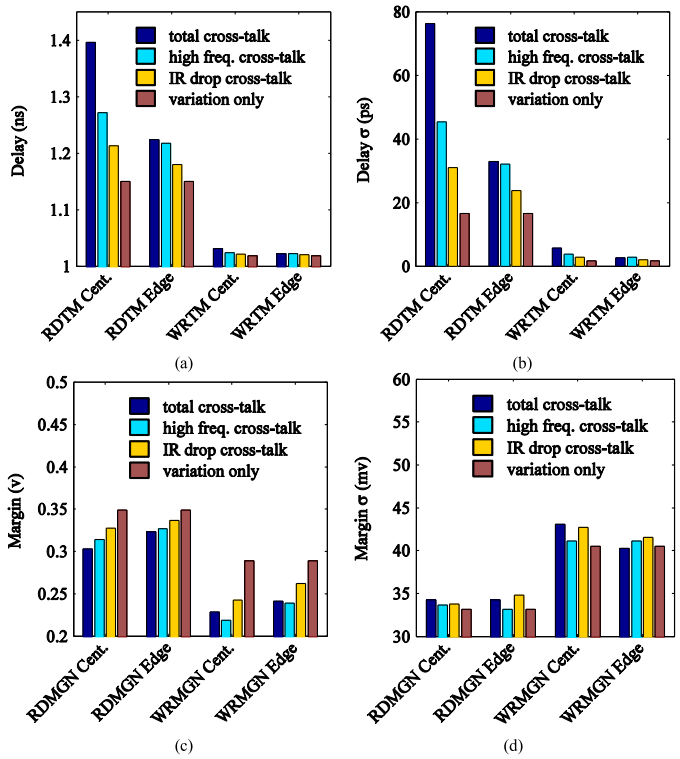


Figure 13: Comparison of the Ldi/dt proximity cross-talk effects including the IR drop cross-talk only, high frequency supply cross-talk, and the combined effect. The IR drop cross-talk captures only the supply variation at DC, the high frequency cross-talk captures the high frequency worst droop, the total cross-talk considers both thermal and high frequency supply variations: (a) the delay metrics of the read time and write time with proximity distance near (vertically aligned with the hotspot) and far (at the chip edge), (b) the standard deviation of the read time and write time, (c) the mean read margin and write margin, and (d) the standard deviation of the read margin and write margin.

2.5 Simulation and Discussion on PDN Configurations

Different die arrangements within stack strongly influence the thermal and electrical conditions within a 3D IC [101]. Two tier stacking branches into two configurations. First, the SRAM tier is further away from the pins without the direct access to the solder bumps but is closer to the heat sink in Figure 14(a). This configuration is referred as the shared PDN with SRAM tier closer to the heat sink, or short for shared PDN w/ SRAM near HS. Second, the stack has shared PDN structure, but the SRAM tier is closer to the solder bumps and is further from the heat sink in Figure 14(b). This configuration is used for all analyses described in Section IV, the shared PDN w/ core near HS configuration. Further, the 3D stack may not necessary share the PDN between tiers in Figure 14(c). Such configuration is the independent PDN with SRAM further from the heat sink, or short for independent PDN. The independent PDN with SRAM closer to the heat sink is not included in our analysis. This is because its electrical property resembles the independent PDN configuration, and the thermal property resembles the shared PDN w/ SRAM near HS configuration. The SRAM parametric metrics of the shared PDN w/ SRAM near HS and the independent PDN are compared against the shared PDN w/ core near HS condition. Both means and standard deviations are reported for all metrics.

2.5.1 Shared PDN with SRAM Closer to the Heat Sink

In this configuration, the shared gird contains 4608 power and ground TSVs and 288 uniformly distributed C4 bumps shared between the SRAM tier and core tier. The active hotspot of size 18 mm^2 dissipates 36 watts of power at the center on the core tier. The thermal cross-talk and supply cross-talk map the SRAM parametric behaviors. Figure 15(a) includes the electrical profile for the shared PDN w/ SRAM near HS case and Figure 15(b) includes the thermal profile counterpart. Both figures include the Shared PDN w/ CORE near HS design for comparison. This analysis

looks at the thermal and supply conditions of a SRAM cell vertically aligned to the hotspot source. These profiles determine the SRAM operating condition accordingly. The shared PDN w/ SRAM near HS configuration has similar supply coupling as the shared PDN w/ core near HS condition, but it's has a lower temperature because the SRAM tier has direct access to the heat sink. A better thermal condition reduces the thermal related variation on the SRAM tier, but the core tier now has a higher temperature. When the power is concentrated, the thermal condition on the core tier is significantly worse than the shared PDN w/ core near HS configuration as illustrated in Figure 15(b). The shared PDN w/ SRAM near HS configuration reduces the read delay by 6% comparing against the shared PDN w/ core near HS case in Figure 16(a). The thermal coupling is more localized comparing to the electrical coupling, and the supply cross-talk dominates the SRAM parametric failures far from the power source. Hence, the differences at the chip edge are less pronounced (less than 1%). The degradation of the shared PDN w/ SRAM near HS configuration is relatively moderate, showing 8% more read time comparing with the baseline (i.e. considering only process variation and no supply drop). The worst case spread deviates 125% from the baseline, as opposed to 200% under the shared PDN w/ core near HS analysis.

2.5.2 Independent PDN

The independent PDN design contains 264 uniformly distributed C4 bumps connecting core tier through dedicate TSVs. The SRAM tier that is further from the heat sink and receives 32 dedicated C4 bumps. The distribution of pins weight the power requirements on each tier and assign 10% pins to the SRAM. Figure 17(a) illustrates the electrical profile for the independent PDN case and Figure 17(b) illustrates the thermal profile along with the shared PDN w/ core near HS design. The independent PDN configuration has similar thermal coupling as the shared PDN w/ core near HS condition, but the supply coupling for the SRAM tier is less compared to the other

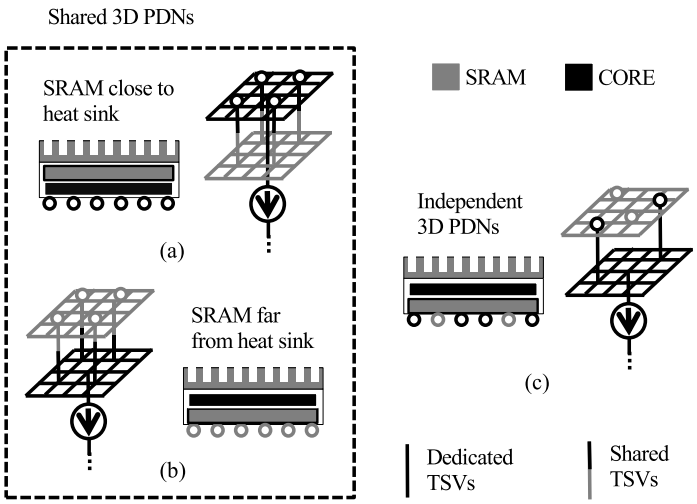


Figure 14: The symbolic tier orders and PDN designs. Three configurations are included in this figure (a) the shared 3D PDN with SRAM close to the heat sink, (b) the shared 3D PDN case with SRAM far from the heat sink, and (c) the independent 3D PDN case with SRAM far from the heat sink.

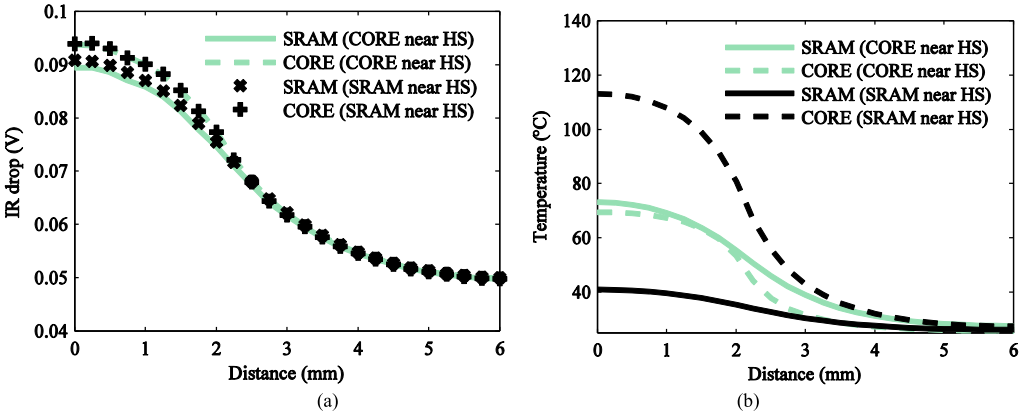


Figure 15: Representative IR drop and temperature variation along the planar distance to the hotspot source considering the tier order. A non-uniformed 36 watts hotspot power of 18 mm^2 is applied to the center of the shared PDN w/ SRAM near HS design: (a) the IR drop observation (b) the temperature observation. The shared PDN w/ core near HS design is included for comparison.

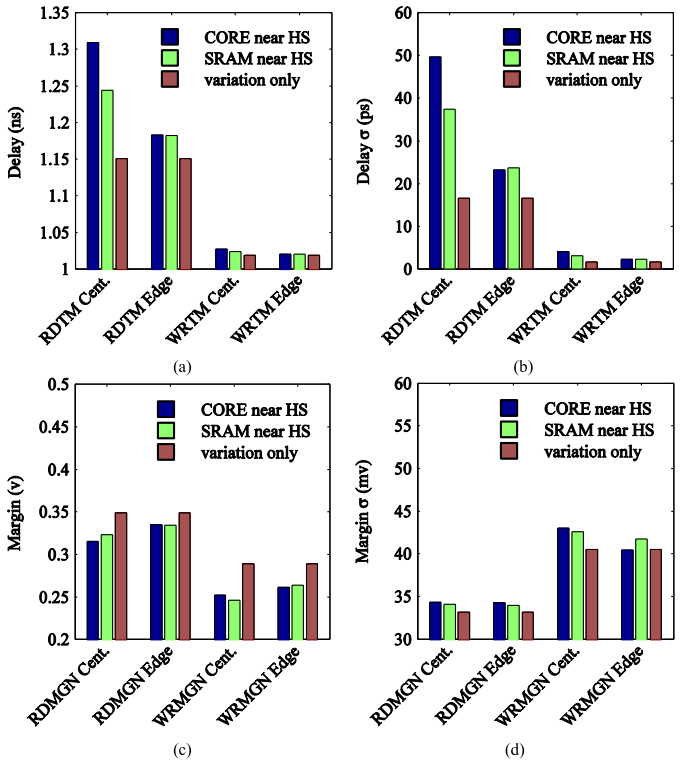


Figure 16: Comparison of the cross-talk effects due to tier arrangements; comparing the shared PDN with SRAM near the heat sink and SRAM far from the heat sink: (a) the delay metrics of the read time and write time with proximity distance near (horizontal aligned with the hotspot) and far (at the chip edge), (b) the standard deviation of the read time and write time, (c) the mean read margin and write margin, and (d) the standard deviation of the read margin and write margin.

two configurations. The independent PDN design sacrifices the supply stability on the core tier but improves stability of the SRAM tier. The electrical condition on the core tier is significantly worse than the shared PDN w/ core near HS stacking configuration in Figure 17(a). The independent PDN have no supply cross-talk, suggesting the SRAM far from the hotspot source will have minor level of thermal cross-talk. The SRAM parametric metrics in Figure 17 confirm this assumption; the mean cell metrics at the chip edge are within 1-2% error of what processes variation predicted. Their corresponding standard deviations are within 15% of the process variation analysis. For SRAM cells aligned to the hotspot source, the results are still better than the shared PDN configurations. This observation is most obvious in the write margin (<1% mean shift), which is a weaker function to the thermal coupling in Figure 17(c). Even at the center of the coupling source, the independent grid structure increases the read time by 7% and its deviation by 75% compare to the baseline variations, the degradation are approximately half of what we observed in the shared PDN w/ core near HS stacking configuration. The benefit of sharing PDN across TSVs is that the horizontal grid impedance effectively reduced due to the extra metal tracks from the SRAM tier. Our independent PDN analysis assumes that the total number of power pins available to the chip (the sum of power pins to the cores and the SRAMs) is constant and these P/G pins are assigned to the core tier and the SRAM tier according to their power demands. For example, if the cores consume 90% of the chip power, 90% of the P/G pins are assigned to the core tier in the independent PDN design. The (marginal) pin reduction combined with lesser horizontal metal tracks adversely impact the core-tier supply stability in the independent PDN design. In order to achieve the same core-tier supply impedance for both shared and independent PDN without using additional P/G pins, the designers may (i) allocate more pins to the core tier, (ii) introduce finer PDN mesh, (iii) introduce extra metal layers for better conductivity, or (iv) introduce more on-chip supply regulation. These solutions are

also bounded by various practical limitations. Increasing pin count (i) is bounded by the available pins provided by a given packaging technology. Introducing finer mesh (ii) is bounded by the design rules within a given technology. Introducing extra metal layer (iii) may adversely affect yield, incur additional lithography masking cost, and complicate fabrication steps. Introducing on-chip supply regulation (iv) requires design effort in additional regulators and on-chip decoupling capacitors. If the supply impedance of the core tier is maintained constant between the shared PDN and the independent PDN, the tier-to-tier supply cross-talk may be avoided. However, the tier-to-tier thermal cross-talk remains an important factor to consider.

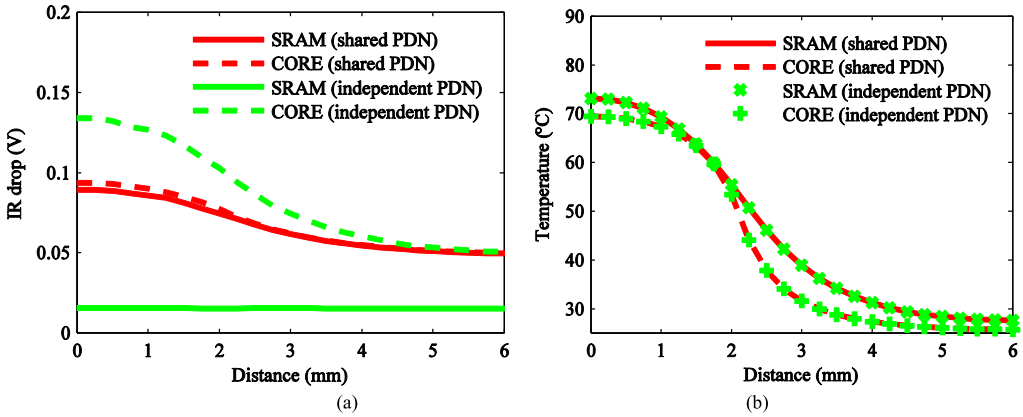


Figure 17: Representative IR drop and temperature variation along the planar distance to the hotspot source considering independent PDN networks. A non-uniformed 36 watts hotspot power of 18 mm^2 is applied to the center of the independent PDN design: (a) the IR drop observation (b) the temperature observation. The shared PDN w/ core near HS design is included for comparison.

2.6 Summary

We have developed a thermal and supply cross-talks aware performance and robustness analysis methodology for electrical-thermal interaction and applied the methodology to study the 3D processor memory stack. Our method considers process variation in transistors, thermal field of SRAM tier, power supply variation in the SRAM tier, and the temperature dependency of wire and TSV resistances. The evaluation

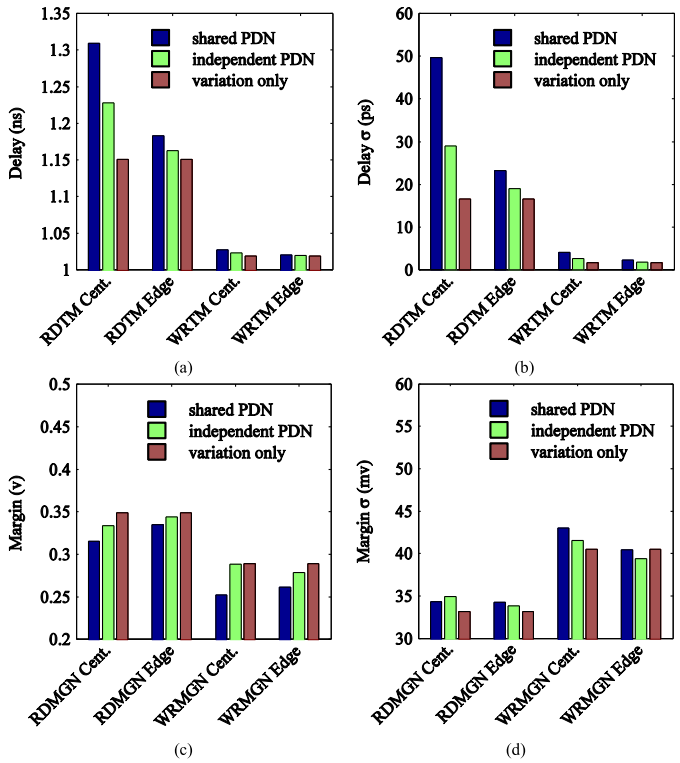


Figure 18: Comparison of the cross-talk effects due to the PDN configurations; comparing the independent PDN and shared PDN with SRAM far from the heat sink: (a) the delay metrics of the read time and write time with proximity distance near (vertically aligned with the hotspot) and far (at the chip edge), (b) the standard deviation of the read time and write time, (c) the mean read margin and write margin, and (d) the standard deviation of the read margin and write margin.

shows that the inter-tier supply and thermal cross-talks may adversely impact the SRAM performance and robustness within a processor-memory stack. Therefore, we conclude that while designing a 3D processor-memory stack, the performance and robustness (i.e. parametric failures) of the SRAM array should consider the power dissipation of the processor. Moreover, the thermal cross-talk changes due to the spatiotemporally varying core power and leads to a time varying hot spot radius and intensity. The electrical characteristics of the stacked SRAM will also have corresponding spatiotemporal variations. Our framework identifies these performance and robustness limitations through modeling the thermal field and the power supply conditions. Our framework identifies thermal cross-talk and supply cross-talk imposed performance limitations and motivates future exploration to adaptive 3D architectures. This framework may guide a holistic solution to adaptive processor-memory power management for a better 3D system performance and tier-to-tier cross-talk resilient design.

CHAPTER III

MODELING ELECTRICAL-THERMAL INTERACTION IN 2.5D PACKAGES - APPLICATION TO EDRAM

3.1 Introduction

The chapter continue the discussion in Chapter 2 on the subject of in-package electrical-thermal modeling. The work in this chapter focuses on the interaction of the interposer based 2.5 D integration. In a 2.5 D environment. The power supply network does not share the same power grid and each die has enough surface area to access power bumps to the interposer and heat sink. The integration increases signal interconnect signal I/O latency with the benefit of lesser die-to-die coupling in power delivery network and temperature. However, even though the coupling is low, the potential thermal and IR drop gradient may still affect the coupled victim circuits differently across the die. The background (coupled) power delivery network skew and temperature may still disrupt device function in an uneven distribution. We model the crosstalk and applies to advanced embedded dynamic random access memories (EDRAMs) for parametric analysis.

The EDRAM has made its way into the commercial micro-processor as the last level cache in-package as in Figure 19 [46]. The traditional on-board DRAM has been assimilated onto the advanced system-in-package (SiP) systems to improve density, capacity, performance, and power. For EDRAMs, the additional reliability precaution is necessary because effects such as thermal coupling affects data retention. In the traditional six-transistor static random access memory (6T SRAM), the cell's retention is not of a concern given a sufficient V_{CCMIN} to ensure greater-than-unity-gain feedback. Unfortunately, by design, the EDRAM cell content expires because the lack

of an equivalent static retention mechanism. The cell content refresh is a constant upkeep to maintain the memory states. Further, the EDRAM cell's availability directly correlates with the refresh rate and the cell temperature. The dynamic behavior in the EDRAM cells makes the physical condition modeling more critical for the SiP integration. Similar to the SRAM cell transistors in a technology, the EDRAM design often defines the minimal device feature size in a process technology. Although EDRAM may be build on a completely different process technology to increase cell retention capacitance [18]. This work focuses on a CMOS process technology without the deep trench capacitor or metal-insulator-metal (MIM) capacitor device support. This defines the scope of the work to allow fair performance comparison between the EDRAM and the SRAM with the same device physics.

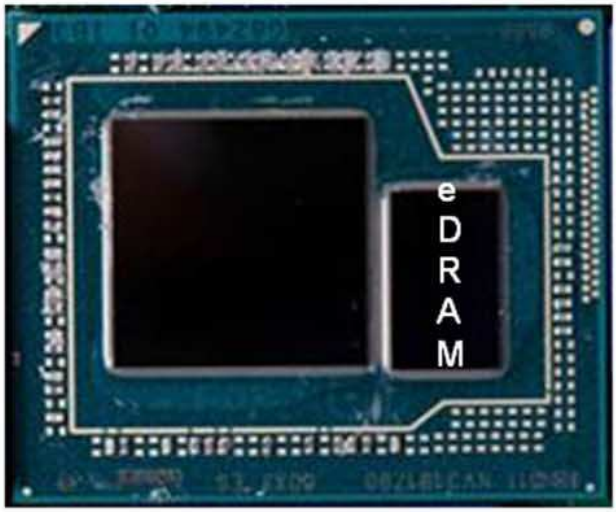


Figure 19: The Haswell CPU-eDRAM MCP package [46].

In order to study the multi-physics-influence on the advanced systems, a memory simulation framework is constructed to evaluate EDRAM memory designs under transient physics conditions. This branch of work builds the multi-physics modeling framework bottom-up and evaluates the influence with the multi-physics interaction to the system reliability. We developed a cross-talk aware performance and robustness

analysis for the SiP to model and analyze this phenomenon. The evaluation framework extend our existing 3D integrated SRAM evaluation framework and include the capability to model EDRAM with multi-physics coupling.

The rest of the chapters are organized as follows: Section 3.2 discusses the related work and contributions of this work; Section 3.3 presents the simulation and analysis framework and the EDRAM operations describing cell to sub-array timing and retention mechanisms; Section 3.4 presents the performance gap between EDRAM cell and 6T SRAM simulation, and how are they closing as the device technology moving from metal-gate devices to finfet devices; Section 3.5 presents different system integrations and results on the EDRAM coupling in these packages; and Section 3.6 presents the chapter summary.

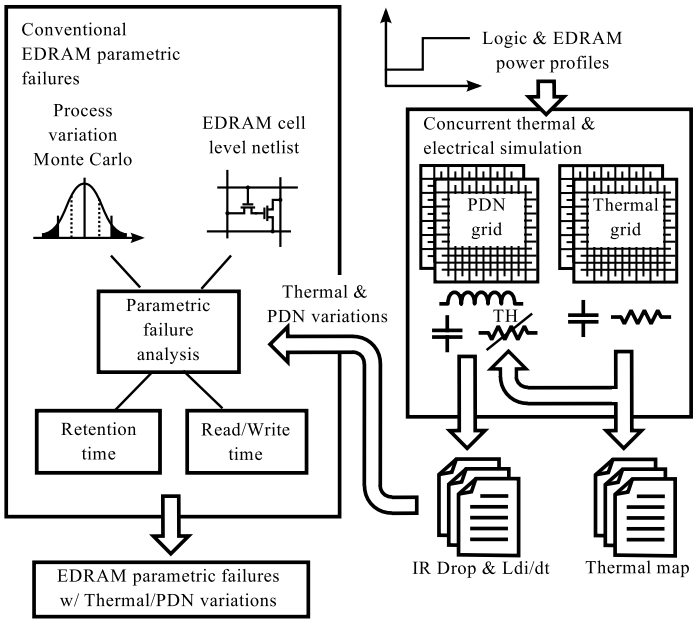


Figure 20: The simulation methodology for thermal and supply cross-talk aware EDRAM analysis. The methodology co-simulates supply and thermal grids with EDRAM analysis.

3.2 Related Work

The 2.5 D interposer integration moves on-board memory into package and improves memory bandwidth. The commercial high bandwidth (HBM) memory has been commercially available in package to improve system bandwidth [48, 29, 43]. The close integration improve I/O power and bandwidth through shorter on interposer interconnects. However, the in-package design suggests the thermal condition of the cell will be impacted by the coupling from the neighboring die. While the existing HBM system are designed with traditional DRAM to replace the last level cache, the more performance oriented gain cell (GC) EDRAM may also applies with the same in package configuration for high bandwidth application [79]. The gain cell EDRAM system trading area efficiency for access time that is on-par with the 6T SRAM system [15]. In order to quantify the thermal coupling and supply condition for in-package to EDRAM performance, the EDRAM's operating conditions should be modeled accordingly.

The embedded dynamic random access memory (EDRAM) is emerging as a promising alternative to the mainstream 6T SRAM design [79, 15, 61]. The EDRAM cells are more compact than SRAM due to fewer transistors in cell design; reducing require cell area by 52 % to 78 % [14]. The associating leakage power also reduced due to device reduction. The read and write decoupled gain cell EDRAM design has high array bandwidth, and may read the cell content without the additional time and energy consuming write-back step [14, 52, 51]. However, the draw back of GC is that the cell logic charge is held in a relative low capacitance storage node (mos-cap or diode) which requires a more frequent refresh cycle. The cell retention for the EDRAM is hence one of the most important parametric variation in design.

There are many experimental work on interposer based 2.5D EDRAM systems [48, 29, 43]. However, most of the emphasis are on the memory bandwidth improvement and not on die-to-die coupling. The experimental system also restrict the system

with defined aggressor core not a generic simulation framework for future systems. This chapter focuses on exploring in-package cross-talk on the EDRAM stability considering the coupled impact of thermal and supply interactions and claim novelty in the modeling effort.

3.3 In-package Memory Analysis Framework

3.3.1 Thermal Modeling with 2.5D Distributed RC Grid

The in-package memory modeling framework models EDRAM robustness and performance in a 2.5 D memory integration. Since the power dissipation in cores is much higher than the EDRAM's, we consider the core die as the aggressor and EDRAM die as the victim. Our simulation framework estimates the temperature on the EDRAM tier due to power dissipation in the processor cores, and evaluates the EDRAM stability. A high level simulation flow is shown in Figure 20. Thermal framework transforms thermal components into an equivalent distributed RC model and uses HSPICE as the backend simulator in Figure 21(a). The resistances of wires in the PDN network are modeled as voltage controlled resistances in the thermal and PDN simulations. Hence, in the simulation framework, the thermal components co-simulates with the PDN components and produce the thermal feedback directly to the planar PDN meshes during the simulation in Figure 21(b). The voltage droop affects the EDRAM's robustness under threshold voltage variations. We consider read time, write time, and retention time as the metrics for EDRAM robustness. The spatial coordinate on each node maps the supply voltages and operating temperatures to the locations in the aggressor cores and the victim EDRAM blocks.

Note that the thermal resistance and capacitance measurements may change due to imperfect attachment during manufacturing such as partial voiding and delamination [45]. Therefore, there may be chip-to-chip variation in the transient thermal properties between packaged ICs. Second, to understand the impact of temperature

variations on ICs, it is important to accurately characterize the interactions between transistor properties and time-varying temperature patterns. In Chapter 4, we discuss further on how to design hardware structure to calibrate simulation models. In Section 4.5.4, an calibration example on coupling improvement and modeling correction has been demonstrated. A combination of modeling and hardware validation refinement ensures the model accuracy. In this chapter, we focus on developing the methodology of modeling itself and separate the validation methodology to the next chapter.

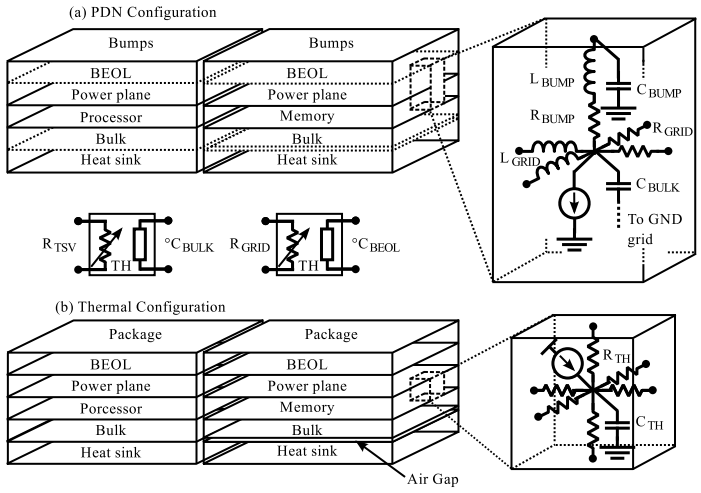


Figure 21: The thermal and supply grid model considers feedback from the thermal effects. The resistances of the PDN grid are coupled to the local temperature (a). The PDN supply unit cell and the corresponding layers it spans are shown on. The thermal grid unit cell and the corresponding layer it represents are shown on (b). The tier PDN resistances are coupled to the thermal BEOL layer and the TSVs are coupled with the thermal bulk layer.

3.3.2 Modeling the EDRAM Parametric Failures

In EDRAM parametric analysis, a high performance gaincell EDRAM cell is considered. A two-transistor (2T) NFET design is modeled for the analysis. The 2T EDRAM design has the highest cell density but requires shielding and pulse read signal for read current limit. While many prior works utilize the PFET for better leakage control and better cell retention [15]. In order to model equivalent speed of

the SRAM, the NFET design is used in Figure 20. The density improvement comes from implementing all devices in the same substrate, and remove the well keep-out area [79]. A similar 3T eDRAM has been implemented 65 nm process in [61]. The related design unlike the proposed 2T NFET EDRAM requires additional read transistor stack, which increases both the cell area and the read delay. The additional transistor also does not improve storage cap to the storage node and is purely inserted for read word-line control with a minor leakage reduction benefit. However, the cell in the prior work enjoys simplified read peripheral circuit design and is relatively simple in the read timing control.

Implementing the proposed 2T NFET design would have additional challenges due to the bitline-to-cell coupling and wordline-to-cell coupling. Because the write circuit turn off by depleting gate charge, the apparent charge in the cell node will be slightly discharged. The coupling effects may be managed through boosted write and pump additional charge to compensate for the coupled charge. Without the boost circuit, the degraded charge reduces drive current to the read NFET and can further couple to the read word line enable across the gate. The read coupling loss may be compensated through boosted write as well. The model intend to build a high performance and high density EDRAM memory system which requires more rigor design margins and lower operation condition fluctuation. The high performance and high density EDRAM cell may form cache banks that are on par with the SRAM cache system in performance.

Pulsed Read: During reading a sneak current path exists through all the unselected cells storing Q=1 in the columns (Figure 22). The sneak current flowing from the unselected RWLs (@Vlogic) creates two challenges: (1) an increased short-circuit power during read, and (2) incorrect sensing due to RBL discharge slowdown. To address this challenge, a pulsed RWL is generated i.e. the selected RWL is enabled for a shorter duration (Figure 22). The sneak current path terminates as soon as the

RWL is disabled. A differential sense-amplifier (with a constant reference) is used per column to sense the small RBL drop during the short pulse. Outputs of the SAs are multiplexed (column decoder) to access one entry at a time.

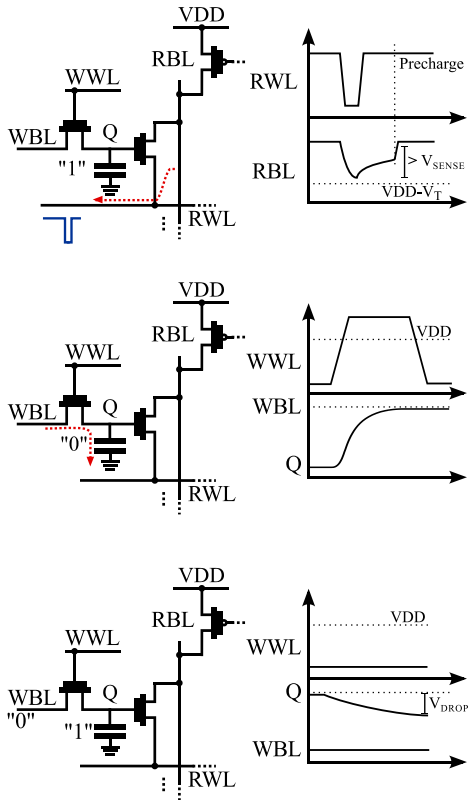


Figure 22: The simulation methodology for thermal and supply cross-talk aware EDRAM analysis. The methodology co-simulates supply and thermal grids with EDRAM analysis.

Other than the cell access, the critical path of a EDRAM sub-bank is very similar to the traditional 6T SRAM system. The flop-to-flop delay is limited by the read wordline driver, cell drive bit-line, sensamp sensing, and read bit-line precharge/sensamp reset. This model assumes the wordline-reset is masked during sense-amp evaluation with a divided-read-bitline multiplexing architecture. Due to the regularity of the EDRAM array, the extracted critical path of the sub-array is deterministic and is defined as:

$$T_{read-cycle} = T_{read-wordline-driver} + T_{cell-drive-bitline} + T_{senseamp} + T_{senseamp-pecharge} \quad (2)$$

$T_{read-wordline-driver}$ is the delay of wordline driver enabling the read wordline to EDRAM access transistors. $T_{cell-drive-bitline}$ is the delay of the EDRAM cell discharge the bit-line at VDD through a pull-down nmos transistors. The temperature-delay relation is simulated from the DRAM's n-mos access discharging 'N' other identical n-mos drain cap on the read-bitline. The actual schematic delay is simulated considering the time required to develop a 100mV of bit-differential, this timing parameter is also critical for EDRAM energy as higher than V_T voltage drop across the bitline may turn on non-selected cells and leaks additionally. The $T_{senseamp}$ and $T_{sensamp-pecharge}$ are related sense-amplifier parameters.

Write Operation: The write using single NFET. Writing Q=0 can be facilitated without modification. For writing Q=1, Operating WWL at $V_{mem} + V_T$, where V_{mem} is the memory voltage overcomes the threshold (V_T) drop in the NFET pass gate and helps writing Q=1. Therefore, using a (V_{mem}) less than the peripheral voltage (V_{logic}=$V_{mem} + V_T$) guarantees a robust write.

The write delay should also be taken into account as the write ties to the cell content update as well as the refresh timing. The write cycle is more simple than the read cycle defined as:

$$T_{write-cycle} = T_{write-wordline-driver} + T_{bitline-drive-cell} \quad (3)$$

$T_{write-wordline-driver}$ is the delay of wordline driver enabling the read wordline to EDRAM access transistors. $T_{bitline-drive-cell}$ is the delay of the write bitline overwrite the cell content through the n-mos write transistor transistors. The temperature-delay relation is simulated from the DRAM's write driver to pump charge to the selected cell as well as discharging 'M' other identical n-mos drain cap on the write-bitline.

Cell Retention: The EDRAM cell retention charge degrade because of the non-regenerative charge inside the cell. The memory cell need constantly refreshed. The cell refresh time reduce the peak throughput and frequency should be minimized to improve system performance. The cell charge is held at Q=1 and the WBL is held at 0 for the simulation in Figure 22.

The cell retention time is measured by the time of fully charged cell discharged to $100mV + VDD/2$.

3.4 EDRAM versus Traditional 6T SRAM Performances

To demonstrate the EDRAM performs competitively to traditional 6T SRAM in cell level. Two flavors of the memory technologies are simulated. The metal-gate 45 nm devices and 16 nm finfet device models are used in the simulation [100, 73]. The corresponding size in related to the minimal feature are shown in Table 5 for SRAM and Table 6 for EDRAM. The sub-array row on the critical path is 32 cell on a single bitline. The read time and write time affect by the supply droop more than the temperature in the given operating region in Figure 23(a,b) and Figure 25(a,b)). Due to low channel doping in finfet, the hotter finfet devices operate faster due to the dominant temperature-voltage-threshold-voltage relationship [17]. The effect does not make significant performance improvement to the finfits at high temperature due to V_T's temperature dependency. However, this suggest the technology dependent correlation should consider the cool corner for analysis to cover the worst case timing, and hot corner for analysis to cover the worst case power. Comparing with the SRAM operation in Figure 25 and in Figure 26, the EDRAM read time is in the same order of magnitude of the SRAM system with the same array configuration with only 8 % read delay penalty. With equivalent configuration, the EDRAM is 60 % slower comparing to the SRAM in metal-gate cell design. The EDRAM write delay are 4 to 5 times worst than the SRAM write delay. Fortunately, because read defines the flop-to-flop

critical path for the memory sub-array access, such performance difference is masked by the longer read access in a random access cycle.

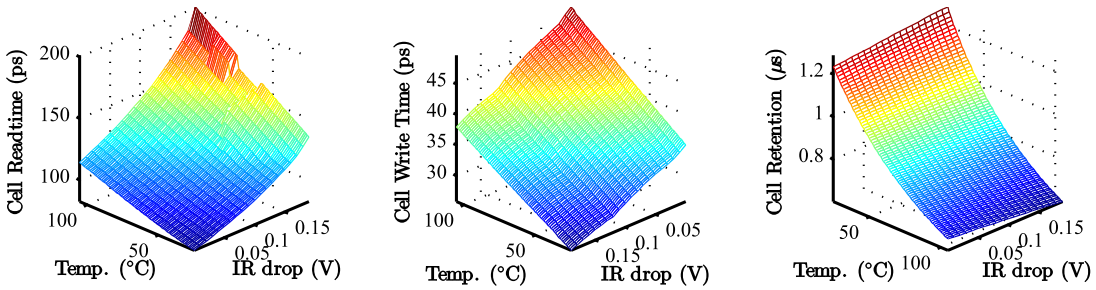
The simulation across the selected two technologies highlight the thermal effect related to leakage in Figure 23(c) and Figure 24(c). The leakage plot for the finfet the design has very low initial leakage, but at a hotter temperature the device leakage increases more rapidly than the metal-gate design. The temperature is therefore more important factor to the EDRAM operation for finfet devices. The retention time, impact the overall EDRAM availability, power, and bandwidth. Quantitatively commenting on the differences at different temperatures, the EDRAM of the finfet cell improve by ˜50% compares to the metal-gate transistor results at the thermal design point (TDP) at 105 °C. Considering the metal-gate devices are four generation older than the finfet devices, the improvement is not a significant. But at the room temperature corner, the retention time for finfet EDRAM is 36 times longer than the metal-gate device. This results suggest that applying advanced cooling and thermal-adaptive refresh for the EDRAM system can extend the system bandwidth drastically and investigating in advanced cooling system and thermal adaptive design are beneficial.

Table 5: SRAM Transistor Ratio

Transistors	Metal-gate 45 nm (ratio)	Finfet 16 nm (ratio)
PULLUP	1	1
ACCESS	1.5	1
PULLDN	2	2

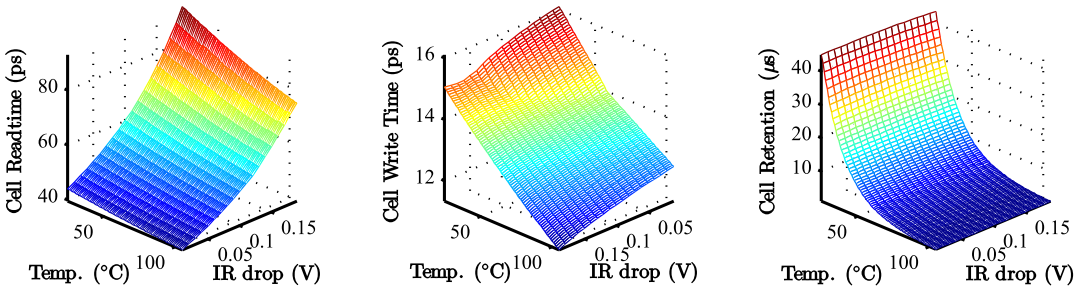
Table 6: EDRAM Transistor Ratio

Transistors	Metal-gate 45 nm (ratio)	Finfet 16 nm (ratio)
WRITE	2	2
READ	1	1



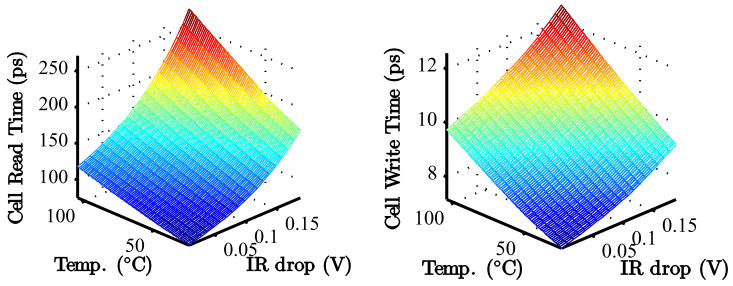
(a) (b) (c)

Figure 23: The delay on metal-gate EDRAM: (a) read time, (b) write time, and (c) cell retention simulation.



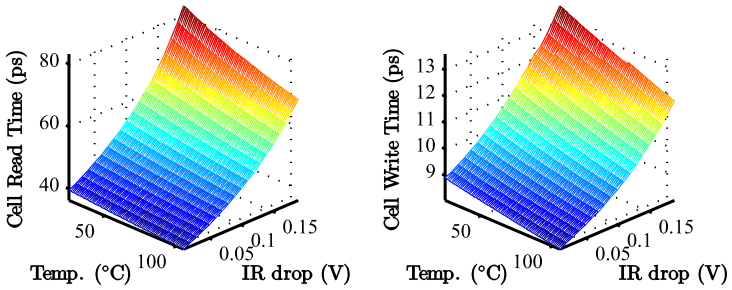
(a) (b) (c)

Figure 24: The delay on finfet EDRAM: (a) read time, (b) write time, and (c) cell retention simulation.



(a) (b)

Figure 25: The delay on metal-gate SRAM: (a) read time, (b) write time simulation.



(a) (b)

Figure 26: The delay on finfet SRAM: (a) read time, (b) write time simulation.

3.5 Simulation and Discussion on Coupling Analysis

3.5.1 2.5D System Configuration

The EDRAM system in an interposer integration will need direct heat sink access in order to achieve efficient heat extraction. However, to achieve a same die thicknesses for different dies in-package is difficult. For a high power system, the processor die is made thicker to make firm attachment to the heat sink. Similar to many passive in-package component, the memory die may not have a direct access to the heat sink. This work models the low power EDRAM system with direct heat sink contact, 200 μm air gap to heat sink, and 1 mm air gap to heat sink configurations. The power distribution for EDRAM is separated from the other logic die but is subjected to thermal coupling on the power supply resistance.

The integrated system with different air gap width on the die may produce different responses. The EDRAM die is modeled with 7 W of uniformed peak power. The coupling CPU is modeled with 72 W of uniformed peak power. The EDRAM and processor are on different power supply plains in Figure 21. The thermal modeling parameters are provided in the Table 7. The power supply network uses the parameters in Table 8.

Table 7: Parameters for Thermal Simulation

Parameters	Thickness (m)	R (W/m • K)	C (J/m^3 • K)
PKG	1 m	20	35.5 K
BULK	100 μ	100	1.75 M
DEV	20 μ	100	1.75 M
BEOL	50 μ	40	4.00 M
BOND	10 μ	100	4.00 M
AIR GAP	0 – 1 m	0.025	1.00 K
SINK	1 m	400	35.5 K

The coupling condition are slightly influenced by the air gap thickness in the modeling. The system thermal coupling from left to right side of the die is captured in Figure 27 and 28. The distance 0 in the figures represent the side of the die thats

coupled to the processor hotspot. The thermal condition is highly skewed to the hot core to the left of the EDRAM die and the overall thermal condition is affected by the thermal properties. The PDN thermal coupling is almost indistinguishable in the modeling due to low thermal gradient across the die. Even though the processor die consume 72 W of power, the thermal coupling is relatively low, the thermal gradient only shift by 5 degrees across the EDRAM die.

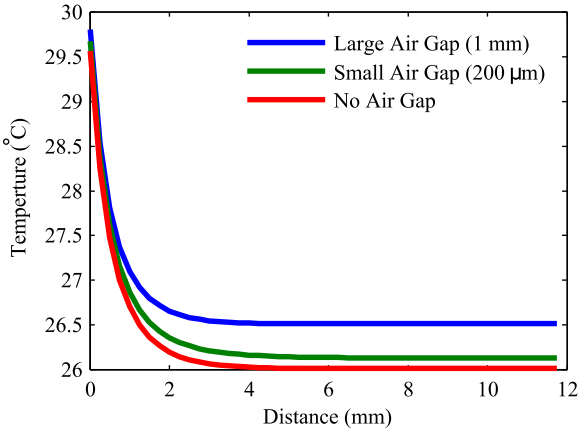


Figure 27: The simulation result for the thermal coupling with a 72 W processor to the left of the die in the package.

The EDRAM under coupling has a slight impact to the system performance. The metal-gate EDRAM exhibit low variation across heat sink gap due to good interposer conductivity. The read, write and retention were not sensitivity to the gap distance up to 200 μm. The gradient across the die is also low, and the timing parameters difference across the die are within 1 % performance gradient. The retention time differs by only 4 % as well. For finfet configuration, the read and write timing in

Table 8: Parameters for PDN Simulation

Parameters	R0 (Ω)	L (H)	C (F)
PCB	94 μ (s)	21 p	240 μ
	166.6 μ (p)		
PKG	1000 μ (s)	120 p	26 μ
	541.5 μ (p)		
BUMP	40 m	72 p	
GRID	28.1 m	3.1 f	93.8 p

Figure 29. Due to the small thermal and voltage gradient, the overall influence to hotspot is lesser than 1 %. The air gap difference also made a very little impact to the EDRAM performance and influenced lesser than 1 %. The refresh time is in Figure 29(b). The read and write delay are affected by the thermal and IR drop by roughly 1 % across the die. The retention time across the die is influenced by the thermal gradient more noticeably and differs by 17 %. This suggest that 2.5D integration design effort is relative simplistic unlike full 3D design. The chip may be design with margin without compromising too much device performance due to parametric failures. However, this advantage comes with the additional interposer connection latency due to lateral distance between the processor die and memory die. This effect has to be taken into account when designing the high bandwidth memory bus.

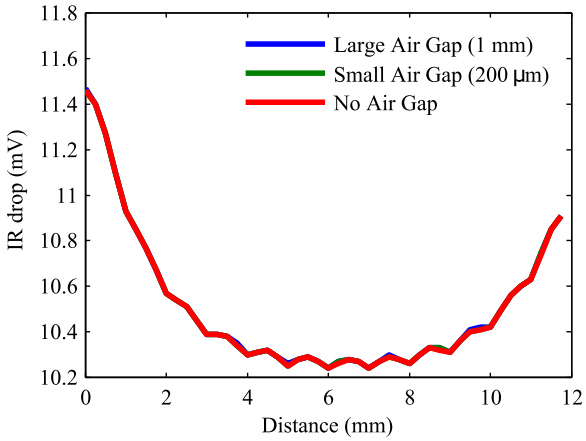


Figure 28: The simulation result for the IR drop due to the thermal modulated power supply network.

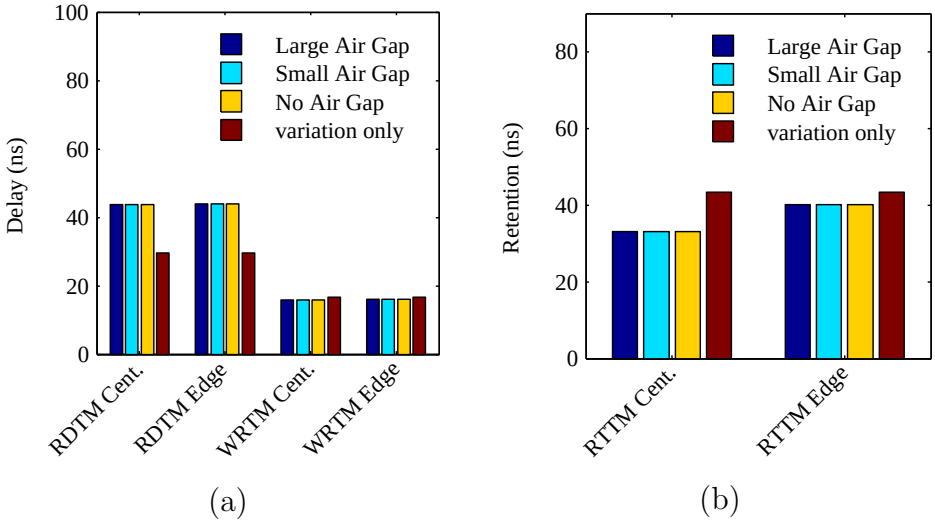


Figure 29: The 2.5 D coupling on finfet SRAM: (a) read time, (b) retention simulation.

3.5.2 3D Stacking Configuration

Different die arrangements within the stack strongly influence the thermal and electrical conditions within a 3D IC [101]. Fortunately, similar to the 6T SRAM memory the power in the stack is relatively low. The EDRAM is still very thermally dependent, and a rational placement for the EDRAM is placing the EDRAM structure closer to the heat sinks. In this configuration, the EDRAM does not have the direct access to the solder bumps but it is closer to the heat sink. The same naming convention used in SRAM is used to identify the EDRAM configuration (shared PDN w/ EDRAM near HS, shared PDN w/ core near HS, independent PDN). The EDRAM cell vertically aligned to a hotspot are evaluated as the worst case response and the location at chip edge is also simulated for minimal coupling condition. The shared PDN w/ EDRAM near HS configuration has similar supply coupling as the shared PDN w/ core near HS condition, but it has a lower temperature because the EDRAM tier has direct access to the heat sink. A better thermal condition reduces the thermal related variation on the EDRAM tier, but the core tier now has a higher temperature and hence, is thermally constrained. The shared PDN w/ EDRAM near HS configuration reduces the read delay by 12 % comparing against the shared PDN w/ core near HS case (Figure 30.a). The thermal coupling is more localized comparing to the electrical coupling, and the supply cross-talk dominates the EDRAM parametric failures far from the power source. Hence, the differences at the chip edge are less pronounced. The same trend appears in the retention time correlation; cells that are further away from the hotspot have lesser coupling (<1.5 %). The center coupling demonstrates the importance of cell retention to cooling. Since the EDRAM is affected by the thermal coupling, placing the EDRAM at the closer-to-heat-sink tier reduce overall leakage and has a center-to-edge retention difference of 34 % comparing to the core near HS case of 24 %.

The independent PDN configuration has similar thermal coupling as the shared

PDN w/ core near HS condition, but the supply coupling for the EDRAM tier is eliminated. The independent PDN have no supply cross-talk, suggesting the EDRAM far from the hotspot source will have minor level of cross-talk. The EDRAM parametric metrics in Figure 31 shows coupling induced read and write time variation to be within 15 %. The elimination of the supply coupling also improves robustness of EDRAM cells (i.e. read time, write time, and retention time) aligned to the hotspot source compared to the shared PDN configurations (Fig. 31). The effect on the critical path readtime delay at the coupling center improves by 16 % by designing out the supply coupling.

The finfet devices also follows the ideal exponential leakage trend (Fig. 32). The thermal coupling degrade the EDRAM retention but improve access read and write time. The hotter tier configuration have marginal performance improvement over the the cooler design in read time by 8% and write time by 21 %. However, the retention time reduced exponentially and the difference is at 80 %. The coupling effect in thermal and power supply is overall undesirable to the system performance due to the exponential dependency between leakage and temperature. The variation across the die are 5 % for the read and 19 % for the write. The tier order appears to affect the EDRAM performance on the same order of magnitude as the planar hotspot gradient. However, the retention is affected by over 90 % across the die. This analysis shows the parametric gradient in the planar direction must be managed accordingly to avoid hotspot in finfet based EDRAM.

The independent PDN results are presented in Figure 33. The performance difference due to tier configuration reduced to 4 % and write reduce to 19 %. Across planar direction the read delay gradient change by 6 % and write delay change by 15 %.

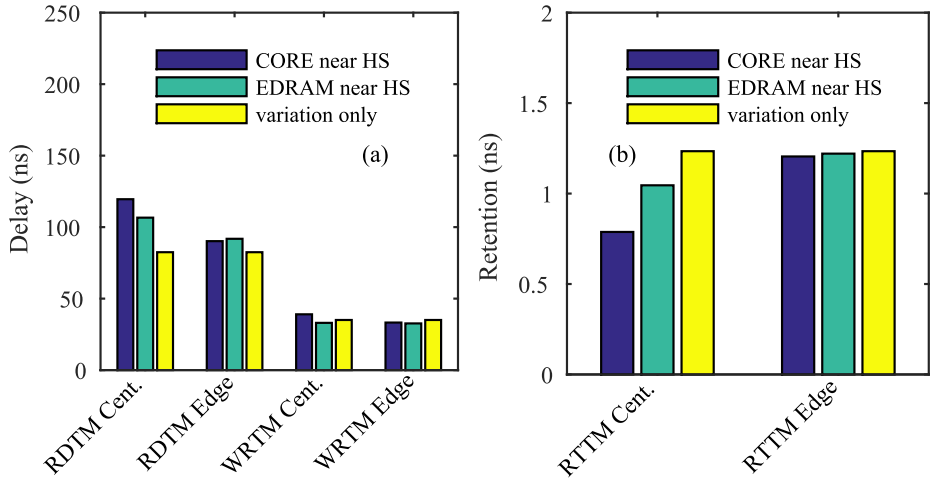


Figure 30: Comparison of the cross-talk effects due to tier arrangements; comparing the shared PDN with EDRAM near the heat sink and EDRAM far from the heat sink: (a) the delay metrics of the read time and write time with proximity distance near (horizontal aligned with the hotspot) and far (at the chip edge), and (b) the mean read margin and write margin.

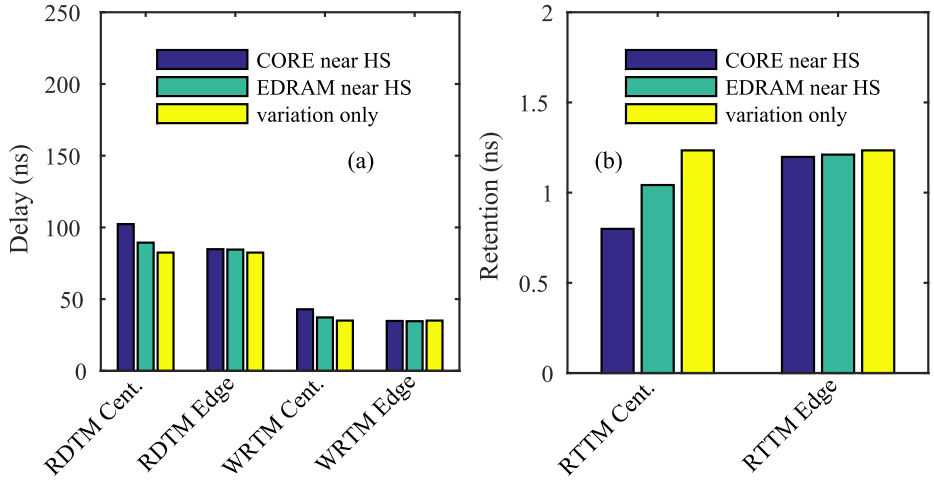


Figure 31: Comparison of the cross-talk effects due to the thermal effect alone; comparing the independent PDN with EDRAM far from the heat sink: (a) the delay metrics of the read time and write time with proximity distance near (vertically aligned with the hotspot), and (b) the mean read margin and write margin.

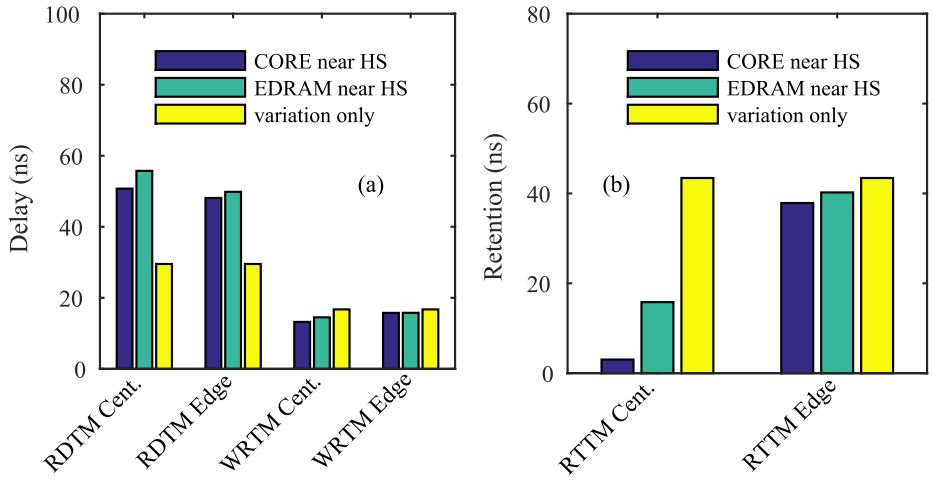


Figure 32: Comparison of the cross-talk effects due to tier arrangements; comparing the shared PDN with EDRAM near the heat sink and EDRAM far from the heat sink: (a) the delay metrics of the read time and write time with proximity distance near (horizontal aligned with the hotspot) and far (at the chip edge), and (b) the mean read margin and write margin.

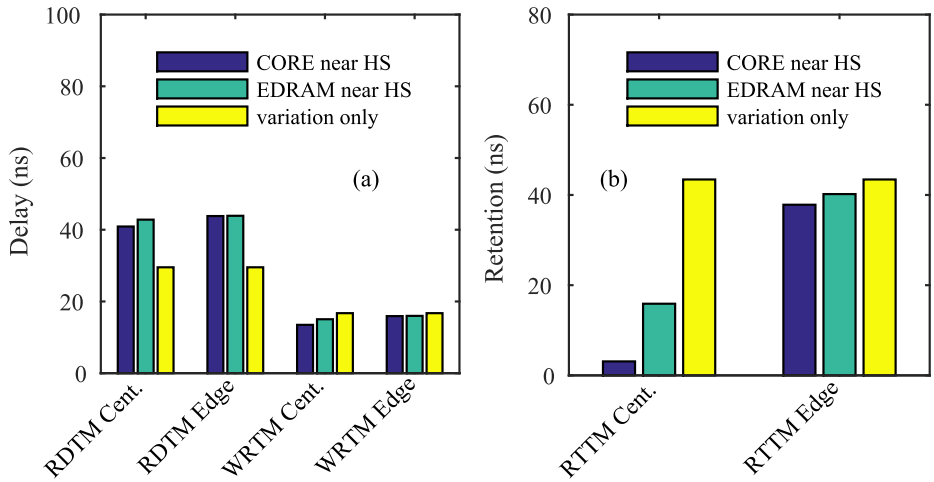


Figure 33: Comparison of the cross-talk effects due to the thermal effect alone; comparing the independent PDN with the EDRAM near the heat sink and EDRAM far from the heat sink: (a) the delay metrics of the read time and write time with proximity distance near (horizontal aligned with the hotspot) and far (at the chip edge), and (b) the mean read margin and write margin.

3.6 Summary

We have developed a thermal and supply cross-talks aware performance and robustness analysis methodology for the gain cell EDRAM in both 2.5D interposer system and 3D processor-memory stack. Our method considers the thermal field of memory tier, power supply variation within the tier, and the temperature dependency of wire and TSV resistances. The evaluation shows that the inter-tier supply and thermal cross-talks may adversely impact the EDRAM performance and robustness within a processor-memory stack. The same coupling effect was also observed in a 2.5D system, but the gradient across the die was lesser than 5 degrees in temperature and two milivolt in IR drop and allows easier design through margin. Further, for multi-die 2.5D integration, the high power die is not creating significant thermal constraints to low power memory die. The coupling and gradient may be designed with margins with lesser than 1 % in die edge-to-edge mismatch. In an interposer integration, a peak power estimate will be sufficient.

The above observations shows that, the horizontal coupling is not of great concern in 2.5D integration, but vertical coupling such as a 3D processormemory stack, the performance and robustness (i.e. parametric failures) of the EDRAM array should consider the power dissipation of the processor. Our framework currently does not consider the presence of advanced cooling techniques for 3D ICs – such as liquid cooling and phase change thermal buffer. With advanced cooling and power delivering techniques, even the 3D integrated system may enjoy significant performance improvement as shown in the 2.5D integration discussion.

The thermal coupling in the memory systems appear to be strongly influenced by the processor electrical-thermal coupling. In the traditional design flow, the maximum activity in the processor defines the thermal ceiling for the package. The memory subsystem is either co-designed with the same thermal constraints on die or integrated on board with only board level coupling. In SiP for both 2.5D and 3D integrations,

the memory and other low power subsystems are integrated to the high performance die from a third party vendor [48]. Because the lack of the spatial and temporal location of the hot-spot information prior to bonding, a transient hotspot stimulation and observation test structure is very valuable to identify coupling in the peripheral low power circuits to model the coupling events.

CHAPTER IV

CHARACTERIZATION OF ELECTRICAL-THERMAL INTERACTION – FIELD PROGRAMMABLE THERMAL EMULATION

4.1 Introduction

The unstructured workload and the corresponding power in advanced multi-core architecture generate spatiotemporally varying temperature. The thermal variation impacts performance and power in the integrated circuits (ICs) [34]. Due to core-to-core thermal coupling, the workload increases core temperature and affects neighboring cores in a multi-core environment. Along with the variable workload, the die-to-die variation in the process parameters and transistor leakage further complicate the predictability of the thermal effect. It has been demonstrated that because the leakage exponentially increases with temperature, the die-to-die variation in chip temperature may be significant, even with the same workload and dynamic power [13]. The thermal condition modulates transistor properties. At higher voltage, digital circuit delay increases with increasing temperature, but at lower voltage, the delay reduces at a higher temperature. Thermal and power management policies should consider this inverse temperature dependency considering transistor properties, circuit types, and process variation [12]. Likewise, there exists a strong correlation between temperature variation and device aging because of bias temperature instability and electrical-migration. To deliver a functional system without excessively over designing the underlie circuits, it is worthwhile to design test structure for thermal excitation and observation circuits just like the common build-in self-test (BIST) module for electrical functional verification and circuit characterization.

This chapter presents a digitally programmable, on-chip, and all-silicon test structure and associated test methodology for post-silicon and on-line characterization of the transient thermal field and its interaction with device properties. This test structure is referred to as the field programmable thermal emulation (FPTE). The proposed structure performs thermal characterization through on-die programmable CMOS based heater array combined with on-die sensors. On die programmable heaters are controlled with integrated registers, to emulate time-varying power patterns and generate time-varying temperature pattern. Multiple digitally programmable FPTE cores are integrated on-chip to characterize the effects of core-to-core thermal coupling. The FPTE cores are augmented with analog temperature sensors to record the temperature patterns, and with digital circuits to characterize the effect of varying temperature on electrical characteristics. The design highlights on-line and "field programmable" characterization framework. A test-chip is designed in 130 nm CMOS to validate the operation of the FPTE and demonstrate its functionality. The measurement results demonstrate the capability of FPTE to generate time-varying and controllable power patterns, sense the resulting temperature patterns, and characterize the performance variations. Multiple (five) FPTE cores are integrated on chip to demonstrate the capability of characterizing core-to-core thermal coupling. Each FPTE core occupies only 0.0375 mm^2 area and dissipates 9 uW of standby power. The demonstration of FPTE validates their application as a thermal test structure designed in conventional digital CMOS process to emulate thermal characteristics of multi-core processor.

The rest of the sections are organized as follows: Section 4.2 discusses the related work and contributions of this work; Section 4.3 presents the emulation and analysis framework; Section 4.4 presents measurement results describing bare FPTE calibration and capabilities; Section 4.5 presents different system integrations and applications utilizing the FPTE framework; and Section 4.6 presents the chapter

summary.

4.2 Related Work

Traditionally, electrical-thermal analysis for ICs has been performed through design-time modeling and simulations (as discussed in chapters 2 and 3). While modeling remains critical, the simulation analysis faces challenges in predicting the run time thermal field in nanometer nodes. The existing transient thermal simulation methods used in fine-grain design-time thermal analysis require accurate estimation of the thermal resistivity and heat capacity of all materials [11, 95, 16]. Many works have studied methods to measure the thermal resistance and capacitance of the thermal interface material (TIM), heat sink, and heat spreader [1, 63, 74, 38]. It has been discussed that, the thermal resistance and capacitance measurements may change due to imperfect attachment during manufacturing such as partial voiding and delamination [45]. Therefore, there may be chip-to-chip variation in the transient thermal properties between packaged ICs. Second, to understand the impact of temperature variations on ICs, it is important to accurately characterize the interactions between transistor properties and time-varying temperature patterns. However, the device parameters experience die-to-die variations. To reduce the design pessimism, the innovations in on-chip characterization circuits have been developed to accurately capture the device properties [65, 78, 3, 72]. Hence, it is important to perform integrated characterization on thermal pattern and device properties for thermal-aware designs.

Considering the post-fabrication and post-packaging variations in the thermal properties of an IC, the on-line characterization of the spatiotemporal variation of the on-chip junction temperature (the transient thermal field) and its impact on the circuit properties are crucial for reliable in-field chip operation [7, 8]. To achieve this goal, one requires on-chip structures (i) for process/device characterization, (ii)

for temperature sensing, and (iii) to generate time-varying thermal field on-chip in a controllable fashion. While significant prior works have described methods for process and temperature sensing, on-chips structures to generate controllable time-varying thermal field is still a challenge.

4.2.1 Designs for Cooling Structure Characterization

Conventional on-die test structures utilize platinum film heater for thermal categorization [72, 97]. The purpose of these test-chips was to explore options in novel packaging materials, die integrations, and cooling structures. The thin-film resistive heater/sensor may have excellent thermal range and stability, but lacks sufficient granularity and field programmability for online characterization. A passive heater component requires off-chip test equipment for data collection which precludes their application for online characterization with multiple controllable heat sources.

4.2.2 Designs for On-die Temperature-Device Interaction

As the off-chip measurement complexity increases with number of heat sources, these approaches are also less scalable and integration with a digital CMOS process. In order to consider the interaction between power, temperature, and CMOS performance, new test structures are required for accurate analysis. Earlier dynamic thermal characterization has been introduced by Poppe, Benedek, Tarter, Reda *et al.* [63, 65, 78, 6]. These designs integrate silicon sensor and heater in tiles for on-chip thermal characterizations. In the analog domain, Altet, *et al.* proposed testing methodology for dynamic thermal effect on device mismatch [3]. These designs mainly aim for characterization during burn-in instead of on-line emulation. The on-chip controllers in earlier works were stateless or singled-state.

4.2.3 Designs for On-die Hotspot Identification

Recent works in the area have shown interest in field-programmable gate array (FPGA) thermal modeling [57, 90]. The in-stock FPGA may be configured as high fidelity thermal sensors, but the emulated heaters may not be effective hotspot sources. Moreover, the FPGA based approaches lack external leakage control mechanism for fine grain electrical-thermal categorization. The contribution of this work is the design of a fully digitally controlled test structure for emulation of spatiotemporal variations in the thermal field considering workload dependent power patterns.

4.3 Field Programmable Thermal Emulator

Parallel to the field programmable gate array (FPGA) in functional emulation, FPTE emulates the thermal field using programmable on-die CMOS based heater array and on-die sensors for temperature and circuit properties. On die programmable heaters are controlled with the integrated registers, to emulate time-varying power patterns and generate time-varying temperature pattern. Multiple digitally programmable FPTE cores are integrated on-chip to characterize the thermal effects in multi-core processing including core-to-core thermal coupling. The FPTE cores are augmented with the analog temperature sensors to record the temperature patterns. The digital circuit characterizes the electrical characteristics under time varying thermal coupling. The conceptual diagram in Figure 34 shows an FPTE chip integrated in a board. The cartoon highlights an on-line field programmable thermal characterization framework on board. A simplified FPTE test-chip to demonstrate the design concept is presented in Figure 35. The proposed design contains following major blocks: SPI registers, digital heater, temperature sensor, and delay sensor.

4.3.1 Heater Hotpot Implementation

The FPTE hotspots at the corners are 600 μm apart horizontally and 750 μm apart vertically. An additional hotspot is placed at the center of the chip for coupling analysis and calibration. The core of the FPTE block is the programmable CMOS heater based on n-well resistor to generate heat within the silicon and near the junction. Each resistor is controlled by an NMOS transistor with binary ('high'and 'low' states to control current through the resistor in Figure 36(a). The resistor bank is arranged in groups of resistors with maximum equivalent resistance as 250 ohm. Each resistor bank is form by 4 x 4 sub-bank tiles. Each tiles are 60 μm x 50 μm in size and shares the control signal to improve intr-bank uniformity. The same dimension resistors are grouped together to form 125 ohm, 62.5 ohm, and 31.25 ohm banks. These 4 banks of resistors are controlled with proportionally sized NMOS transistors to generate 16 power levels with maximum 165 mA per bank at 3.3 V. Hence, the design has an equivalent 16 quantization levels with the maximum instantaneous heater power of 544.5 mW and with the granularity of 34 mW. The total area of the heater hotspot is 250μm x 150μm i.e. 0.0375 mm^2. The voltage of each bank may be controlled with external voltage source in this design (or using on-chip voltage regulators). The internal registers can then tune the individual heater output at a fine-grain level.

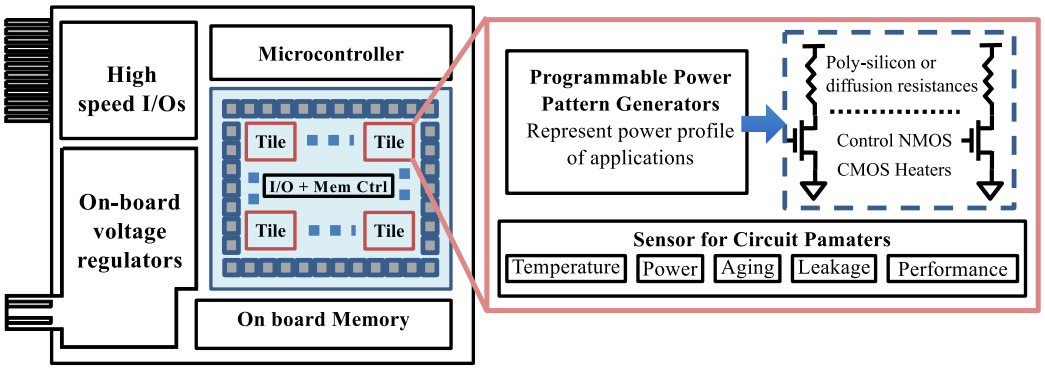


Figure 34: A conceptual diagram of a field programmable thermal emulator (FPTE) integrated in an instrumentation board for thermal characterization.

The latch based registers for heaters and sensors have a footprint of 250μm x 150μm. The fill-factor for a single FPTE is designed to be 50 % and the density may further improve if SRAM is used for the register cells. The heater drivers are designed to sink the full swing current with rise/fall time of 20 ns. Because each resistor has distinct on/off states, the actual resistor layout is sub divided into 16 subbank tiles in Figure 36(b). Smaller subbanked resistor improves the hotspot uniformity within the bank. Within each tile, the resistors are arranged into a common-centroid layout to improve quantization matching. The tiles form two rows with butted NMOS grouped in the center isolation well. The isolation well around the NMOS enables the opportunity to control the leakage through body biasing. This feature enables the opportunity to model and control the device leakage which pure resistive heater may not model readily.

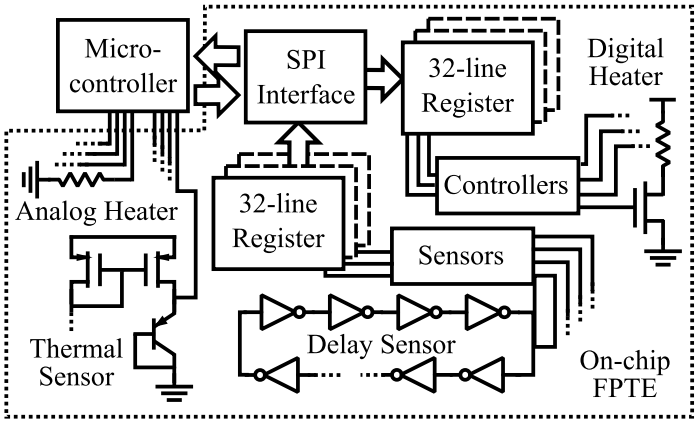


Figure 35: Block diagram of the system contains the external microcontroller interfacing the on chip SPI. The on-chip SPI interface programs heater registers for heating and read data from sensor registers. The microcontroller also applies voltage across analog heaters and senses thermal sensor voltages with built in DAC.

4.3.2 Programmable Register Implementation

To emulate the time-varying power pattern, the on-chip heater register can program up to 32 states in Figure 37. The states wrap around and form a periodic thermal profile when the counter reset every 32 clock steps. The clock pin and enable pin may

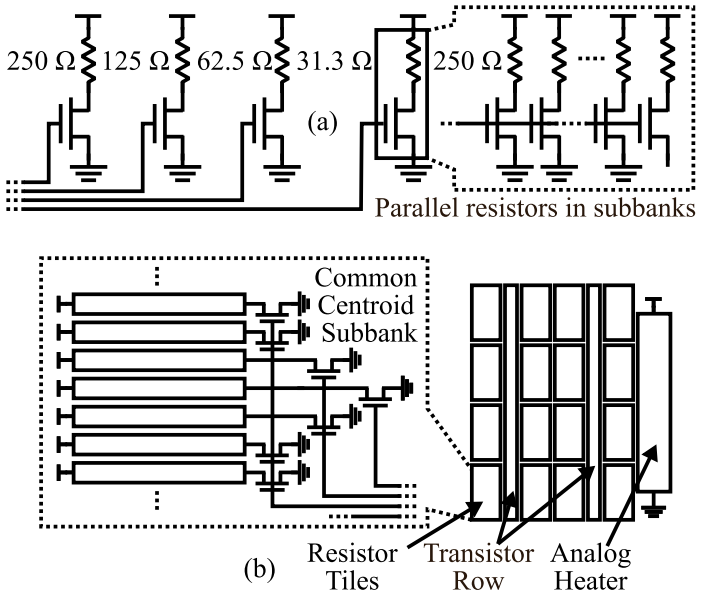


Figure 36: The design of the digital heaters: (a) the circuit schematic of the digital heater and (b) The floorplan of the heater. Binary sized heaters are created with the same sized resistor and transistor pairs to improve regularity. The transistor gates in each group are tied together to form less resistive heater while maintaining the same hotspot density. The heaters are arranged into common-centroid tiles. Smaller tiles ensure power density uniformly (due to quantization) while larger common-centroid groups ensure better matching.

stretch clock to a defined cycle emulating longer steady state response. All heater registers are dual ported so the registers may be updated in the background when the heater applies the power pattern in parallel. The programming can be performed using the on-chip logic cores in a multi-core processor. However, in the absence of a logic core, in this test-chip, an external microcontroller acts as the programming module for the FPTE blocks. We have also included externally controlled heaters in the design to enable background heating, if required, to characterize the sensors without using the digitally programmable heaters. Each background heater is designed to be 40 ohm.

4.3.3 Temperature Sensor Implementation

For sensing temperature, the conventional BJT based analog sensors are designed [13]. The design is shown in Figure 38(a). The outputs of the analog sensors are quantized using external analog-to-digital converters (ADCs). For characterization of the interaction between the delay and transient temperature patterns, digital ring oscillators (performance sensors) are integrated within the FPTE blocks. The delay based sensors store temperature history in-8 bit word and the FIFO holds up to 16 entries in the register in Figure 37. The nine-stage ring oscillators (RO) are designed to update the counter at 500 MHz at nominal 1.2V. The RO counter has 32-bit depth counter, in order to fit the data into 8-bit register we allow external pins to multiplex the counter register into 4 segments in Figure 38(b). When accuracy is important, the microcontroller signals the chip to shift the upper and lower words into multiple register lines while holding the counter value. When accuracy constraints are relaxed, the microcontroller only shifts the upper 8-bits of the counter that contains data. The upper non-zero register position depends on the sampling window, and the count may or may not utilize the upper registers. By monitoring the digital bit string generated from the ROs we can characterize the interactions between temperature

and performance.

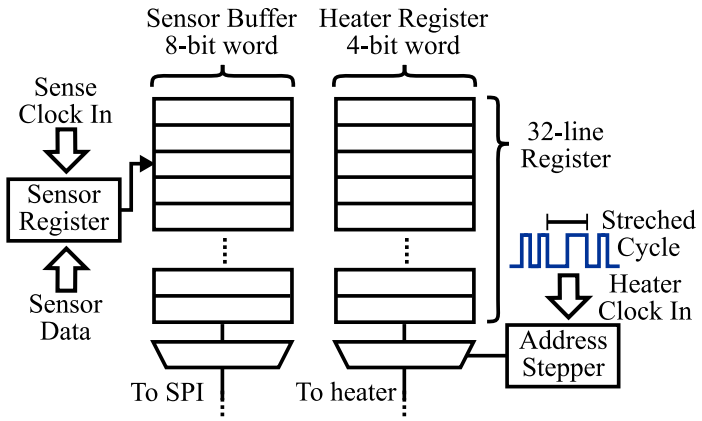


Figure 37: The control logic for the sensor and the digital programmable heater. The sensor is connected to 8-bit word with 16 registers. The heater has 4-bit word with 32 registers. The externally controllable clock may utilize clock enable to stretch clock for finer control of the heater pattern with limited registers. The sensor's full 32-bit counter is multiplexed from 32 bits to 8 bits for shorter register lines. The sensor buffer may be programmed to capture longer bit range by storing the segments into multiple 8-bit registers.

4.4 System Measurement Results

The test chip was designed and fabricated in 130 nm CMOS process. Figure 39 shows the die-photo of the test-chip and Figure 40 shows the layout of the individual FPTE cores. The 2 mm x 1 mm die hosted five FPTE structures sharing the same SPI I/O. The package housing the die was wirebonded ceramic LCC52 by Kyocera. The socket assembly was lidded LCC55 from 3M on a 106.5 mm x 79 mm printed circuit board with passive cooling. The experiment for the test chip was done with the setup in Figure 41. The chip communicated with an external microcontroller directly with digital controls. The heaters directly drew power from the function generators due to the current sourcing limit on the controller. The program stored on the microcontroller orchestrated the heaters and sensors activities. The microcontroller fed control signals from preloaded ROM onto chip's SPI bus and buffered sensor reading for serial port transmission to connected computer. While the chip readings

were completely visible from the microcontroller, we utilized the oscilloscope and PXI interface for data verification and waveform capturing.

The FPTE uses on-chip digital heaters and sensors to emulate time-varying power patterns on-chip, generate the corresponding spatiotemporally varying temperature pattern, and characterize the resulting variations in circuit properties. The application domains of FPTE include a thermal test-vehicle as well as an on-line thermal test-structure.

4.4.1 Steady State Calibration

The resistances of the heaters are measured in Figure 42 to estimate the generated power density at various conditions. Five digital heaters exhibited linear response to external voltage control and digital control. Figure 43 demonstrates the ability program time-varying power pattern in the heaters. We used the on chip digital heaters to program a sine wave and saw tooth wave power pattern. The first half of the figure showed the SPI program of the resistor registers and the second half showed the continuous power patterns when the heater clock enabled. In the sine wave generation, note that the current step up before the sine wave started, this is due to the feature for programmer to write-through register to the heater for on-line pattern generation. The reconstructed waveforms had a period of 250 Hz that was limited by the speed of the external heater clock which in the future may be improved by dedicated function generator.

The on-chip sensors were calibrated in a simple thermally isolated chamber in Figure 44(a). The external heating material heated up the stack assembly until equilibrium and sampled reading from the chip was recorded along with the external thermo-couple module. The temperature inside the chamber slowly decayed while the temperature was used to evaluate the sensors' readouts. The sensor readings from the chip were captured in Figure 44(b) and Figure 44(c). The digital performance

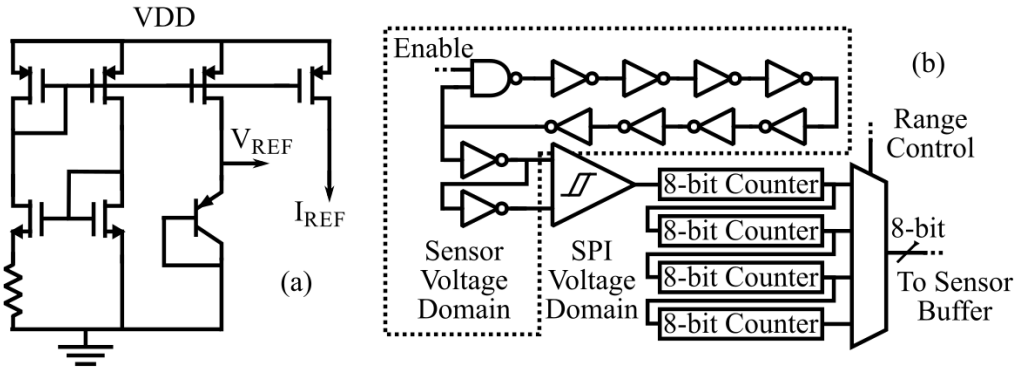
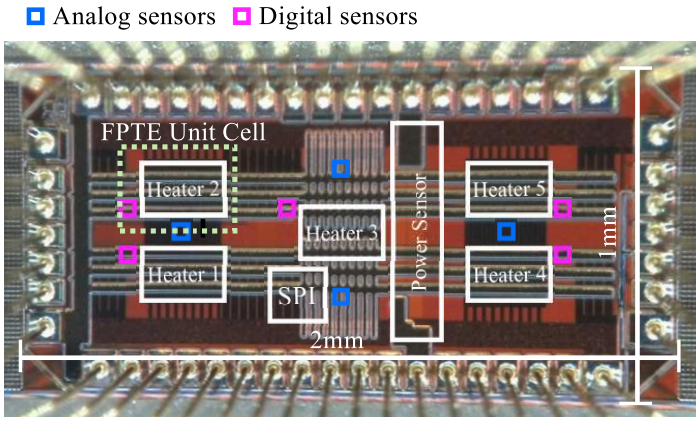


Figure 38: Schematic of the sensors: (a) the analog sensor, and (b) the digital sensor. The analog sensor design is based on the prior design [13]. The sensor output is the VBE of the BJT, labeled VREF in (a). The digital sensor is within a tunable voltage domain and interfaces with the counter through a level converter. The oscillator driven counter has 32-bit range. The output of the counter feeds the 8-bit sensor buffer.



Technology	130 nm CMOS
Voltage	1.2 V and 3.3 V
Area	1 mm x 2 mm
Package	Wire bond LCC52

Figure 39: The die-photo of the test-chip. The chip contains five digital sensors and five digital heaters. They are arranged in a symmetrical design. The analog sensors are placed in north, east, west, south location of the chip for easy access to pins.

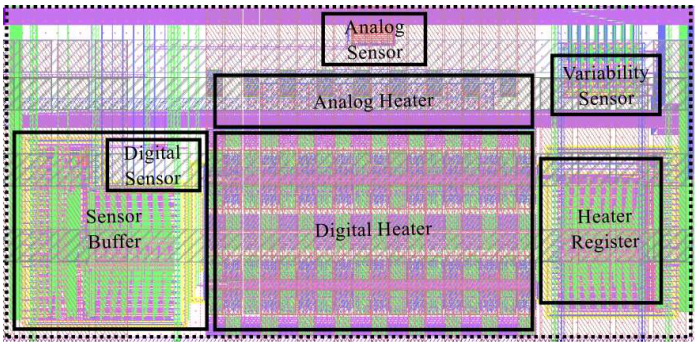


Figure 40: The layout of the FPTE block. The block contains a digital sensor with associated FIFO buffer, a digital heater with its pattern programing registers, a traditional analog temperature sensor, and an analog heater.

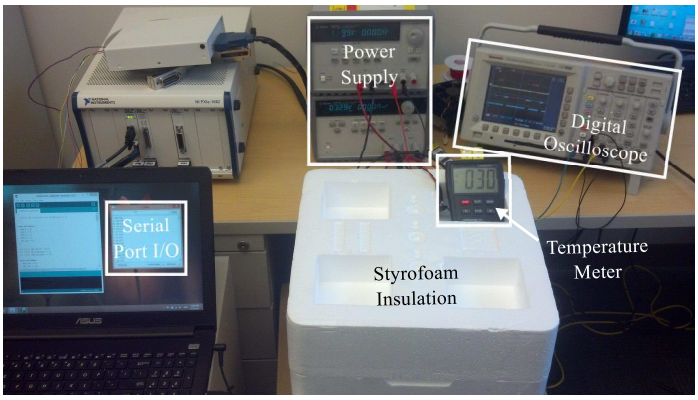


Figure 41: This is the snapshot of the calibration environment. We built a chamber to emulate a closed system for resistor profiling and sensor calibration. We utilized the PCs serial I/O terminal to communicate with the microcontroller and collected data from the chip.

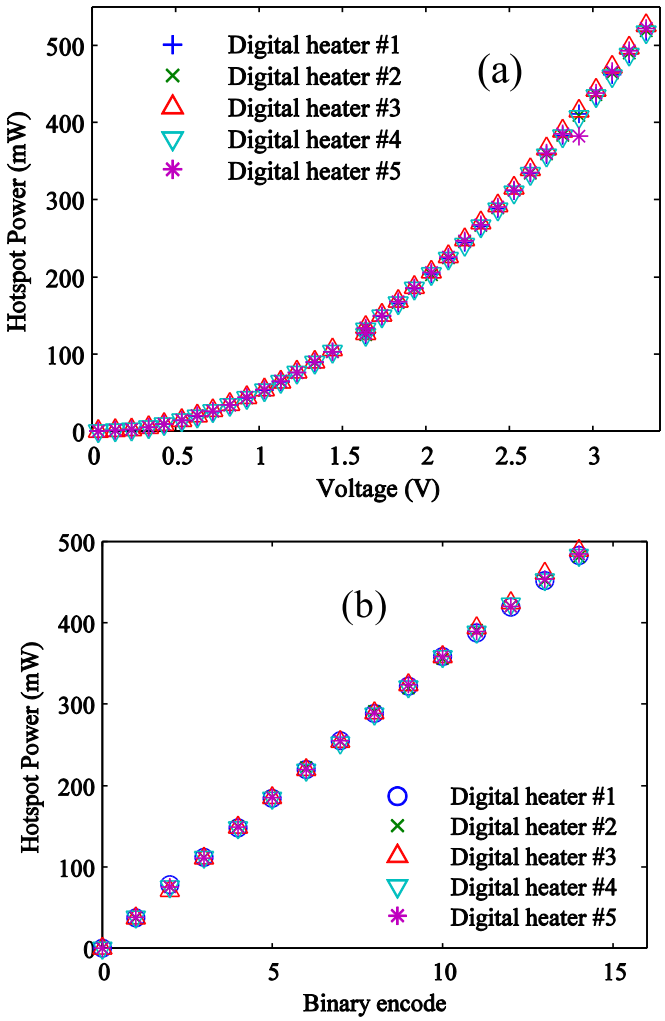


Figure 42: The measurement results showing the DC property of the heater: (a) heater voltage versus generated power density for a given binary code and (b) generated power density versus binary codes for a constant heater voltage. The figure shows linearity of the digital heaters (20 Ω when all resistors on in the bank). The results in (a) show that due to V2/R response of the generated power, the generated power is not linear to the programming voltage. The digital encoding versus current shows high linearity in (b) down to low-power regions. The digital controllable granularity is 34 mW per unit at 3.3 V. The binary weighted resistors match within 2.7% of the theoretical calculation at 3.3 V. Combining the controls in (a) and (b) increase the overall controllability of the heaters.

sensor was sampled for 64 microseconds. The second byte of the counter was captured by the register and had an oscillation frequency 520 MHz. Systematic offset in the digital sensor may be corrected by sampling additional lower byte and perform post processing for calibration.

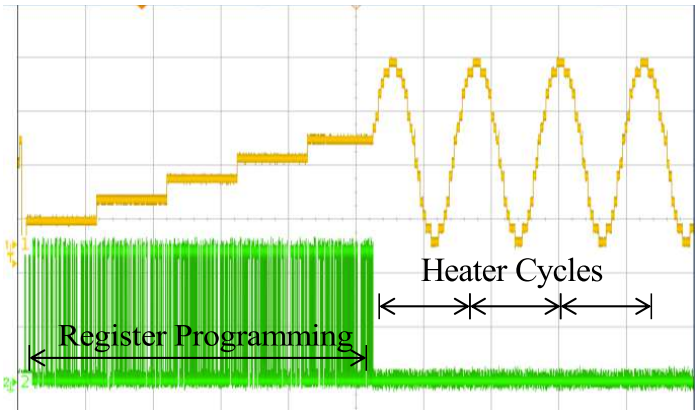


Figure 43: The demonstration of the heater programming method. This figure shows heater banks are programmed to sine wave pattern.

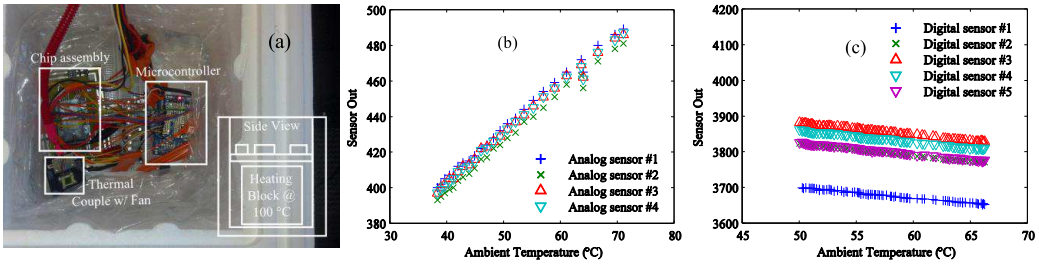


Figure 44: The calibration of the on-chip sensors: (a) the experimental setup for sensor calibration, (b) response of the analog sensor, and (c) response of the digital sensors. Inside the chamber are the microcontroller, chip assembly, and a thermal couple with fan. Under the board is a container holding fluid at 100 degree Celsius releasing heat until equilibrium. Then the microcontroller collects thermal information as the temperature inside the chamber drop steadily. On the microcontroller package there was an LM35DZ sensor for additional temperature recording. Analog sensor reading of the ambient temperature is presented in this figure. This data was used for digital delay calibration. The quantization was done on an external microcontrollers ADC. The ADC reading has a 0.4 °C per unit sensitivity. The calibration of the digital sensor: lower 16-bit from digital sensor versus temperature plot is shown in 10(c). The sampling time was 2 microseconds during the capture phase. The resolution of the sensor is 0.303 °C.

4.4.2 Transient Thermal Emulation

To study the potential of emulating time-varying power patterns, we have applied different power patterns to individual digital heater # 1 and heater # 4. The power waveform and corresponding performance variation (due to the coupling between power, temperature, and delay) are collected with the digital sensor on chip. The capturing window length was constrained by the microcontroller's available memory. When the data filled up the controller RAM, the controller held old the thermal generation and transferred the sensor table to the computer's serial port before continuing. This allowed 128 ms samples with sampling period of 2 ms. Figure 45 and Figure 46 show two different power patterns. Figure 45 is a period power pattern (mixture of sawtooth and sine-wave) and Figure 46 is an arbitrary pattern. Note that the arbitrary power profile may be driven by realistic power trace from measurements of current processor or from architecture simulators for predictive architectures. Within the package we were able to simulated extremely high power density on the test chip. The framework may be coupled with architectural simulation to achieve accurate thermal estimations.

4.5 Illustrative Applications of FPPE

The FPTE uses on-chip digital heaters and sensors to emulate time-varying power patterns on-chip, generate the corresponding spatiotemporally varying temperature pattern, and characterize the resulting variations in circuit properties. The application domains of FPTE include a thermal test-vehicle as well as an on-line thermal test-structure. As a thermal test-vehicle, FPTE has the advantage over existing thin-film heater based approaches due to its compatibility with standard CMOS process, ability to generate controllable and time-varying power patterns, and directly characterize the effect of temperature patterns on device characteristics. As an on-chip test structure, an FPTE block may be embedded within a microprocessor core with

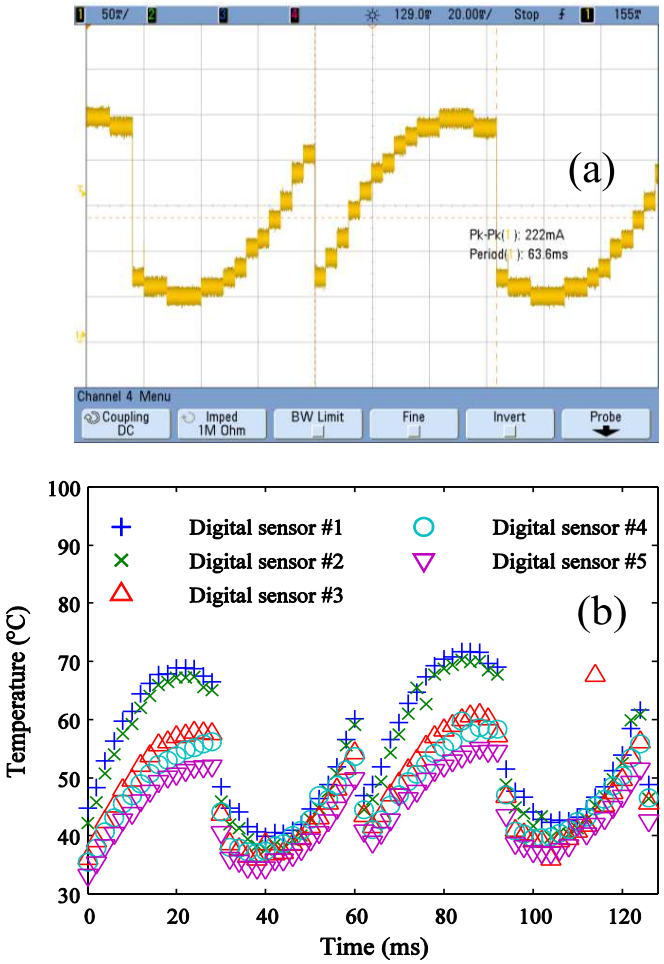


Figure 45: The measurement results showing complex power pattern and associated change in the temperature. (a) the applied power pattern and (b) variation in the sensor output, the calibration from Figure 44 was used to converter the sensed performance to temperature. We applied heater pattern of sine with sawtooth in the same chip on different heaters (# 1 and # 4). The combined effect was observed in the digital sensor output. We may observe the temperature gradient and the transient difference with each digital sensor output. The result demonstrates the ability of the FPTE to characterize the coupling between varying power pattern, temperature, and performance.

minimal overhead, because its low standby power and small area. The FPTE test structure can also serve as an advanced cooling structure benchmark to evaluate the operating thermal field and predict system thermal response with emulated silicon data. In this section, we discuss three applications of FPTE, performed in collaboration with other researchers at Georgia Tech.

4.5.1 Direct Thermal Characterization

As a thermal test-vehicle, FPTE has the advantage over existing thin-film heater based approaches due to its compatibility with standard CMOS process, ability to generate controllable and time-varying power patterns, and directly characterize the effect of temperature patterns on device characteristics. Application of FPTE as a test-vehicle to evaluate advanced microfluidic cooling has been presented by Wan *et al.* [86]. More details on the fabrication of the microfluidics pin-fin and control of the cooling experiments can be found on the PhD thesis of Wan of Georgia Tech. In the experiment, the silicon interposer with etched micro pin-fins was attached directly with the FPTE die in Figure 47. The CMOS chip was attached to the center of the microgap by a thin layer of epoxy which is above the pin fin area. Then the microgap was attached to the back of PCB by tape. There was a small rectangular hole at the center of the PCB which was used to expose the chip. The chip was then wirebonded to the PCB which was soldered and connected to outside circuit. Agilent e3620A dual power supply was used to provide 1.06 V voltage to the four temperature sensors, and variable power input to the heaters. Keithley 2401 sourcemeter was used to provide 3.0 V Vdd and measure the leakage current through the transistors. The sensor output voltages were collected by Agilent 34970A data acquisition unit and converted to temperatures. Two big rectangular holes were cut into the PCB to expose the fluid vias of the microgap. Then two nanoports were placed upon the fluid vias and attached to the microgap by epoxy. The FPTE was used to directly

characterize the effect of fluidic cooling on the circuit properties, for example, the trade-off between flow rate (cooling power) and the potential leakage power saving. The test structure may turn off the CMOS components on chip and utilize the system for leakage measurement. The on-chip temperature may be captured with on chip sensors and the the leakage number may be captured with an external current sensor in Figure 48(a). The active fluidic cooling is relatively effective considering the stack thickness on the chip. In Figure 48(b), from the measurement, the different cooling structure influence the system thermal condition. The observed leakage is different for different cooling methodologies at the same temperature. This non-ideal behavior is slightly disadvantage to the inferior thermal-managed system. This difference is mainly due to the IR drop effect. Because at the same temperature the better-cooled system requires more current to power the chip, the bondwire and package apply a slight IR drop between the supply and ground. The difference made little difference in the total power but is relative noticeable comparing to the leakage power.

Without active thermal management the observed device to ambient thermal resistance is 52.6 K/W. The average thermal resistance of the forced air cooling is 43.6 K/W while that of the microfluidic cooling is 26.9 K/W. While on the chip level the heat flux density is only 50 W/cm^2, the enabled special hotspot analog heater density reaches 1500 W/cm^2 per four 200μm x 20μm stripe heater 40 μm next to the hotspot sensor circuit (# 1,# 4). The uncored sensing circuit (# 2,# 3) are located at 400μm away from the hotspot stripe and is at the corner of the chip. Measuring the temperature difference between the sensors we calculated the thermal resistance between hotspot and uncored circuit is roughly 4.7 K/W for all cooling methods. In the experiment, the chip is externally attached to the microgap, which results in additional thermal resistances from conduction resistance of epoxy, silicon oxide layer, and spreading resistance from epoxy to microgap. The conduction resistances of epoxy and silicon oxide are large due to its low thermal conductivity. In order to

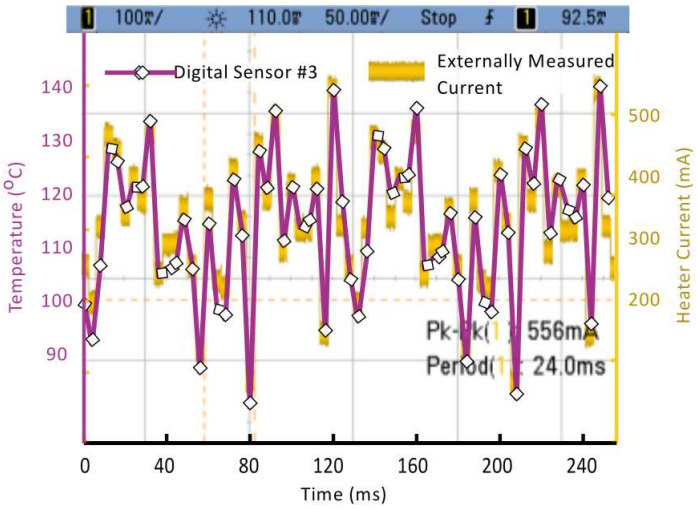


Figure 46: The measurement results showing time-varying arbitrary power pattern: the applied power pattern and variation in the sensor output, the calibration from Figure 44 was used to converter the sensed performance to temperature. We captured the center digital sensors output. The experiment shows the ability of FPTE to generate controllable time-varying arbitrary power pattern. The result demonstrates the ability of the FPTE to characterize the coupling between time-varying power pattern, temperature, and performance.

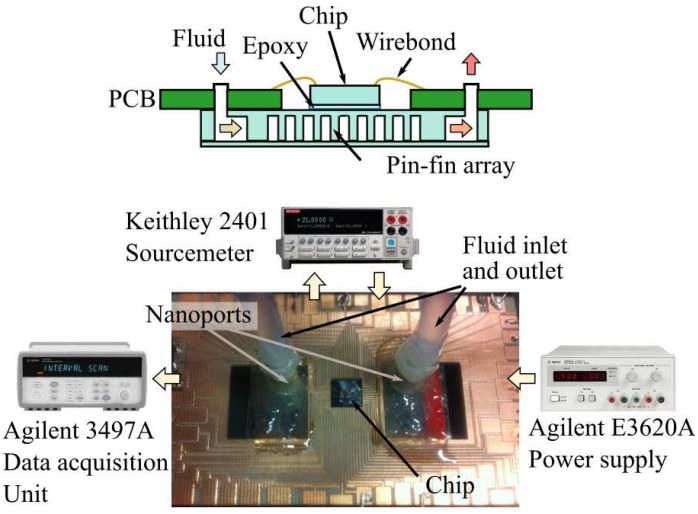


Figure 47: Schematic and experimental assembly of CMOS chip, microgap, PCB.

have effective hotspot diffusivity system, a direct embedded cooling structure should be employed [67].

4.5.2 Characterization of Thermal Coupling

As an on-chip test structure, an FPTE blocks may be embedded within a microprocessor core with minimal overhead, because its low standby power and small area. A built-in self-test routine may apply test power patterns to validate/ensure thermal fidelity of a specific packaged IC considering process variations as well as the time-dependent degradation in the thermal properties [13]. The thermal characteristics of the packaged chips may be extracted in the form of frequency domain thermal filters as presented by Kung *et al.* [44]. More details on the filter extraction methods can be found in the PhD thesis of Jae-Ha Kung of Georgia Tech. The extracted filter was used to accurately predict transient temperature pattern for spatiotemporally varying power patterns. The filter also estimate power pattern from measured temperature accurately, which implies FPTE can facilitate on-line real-time temperature/power prediction in an IC. The thermal filter may be extracted with applying individual heater with periodic heating profile as stimuli and collect per sensor output. The filter is constructed with fast Fourier transform in Figure 49. The phase information may also be captured with the same methodology in Figure 50. The reconstructed filter from the sensor phase array is a powerful tool to predict thermal behavior at position without sensors, which has been demonstrated by Kung *et al.* One limitation of the testchip setup is that the phase between sensor differs only by one to two degrees because of the small die area. This suggest the hotspot information is almost a super-positioned response on an 1 mm x 2 mm die. The observed 3.5 % error drowns our phase detecting feature implemented on chip. In order to truly verify the thermal filtering effect the die size need to be significantly larger. However, the same observation also suggests that the larger dies in the range of 100 to 200 mm^2 are unlikely

to have phase delay exceed one cycle. In practice, simple weighted summation circuit may be sufficient to find the hotspot coupling on chip.

From the filter information in Figure 49 and in Figure 50 the methodology may be used to generalize for architectural or power-gating meta-signal to predict thermal behavior at arbitrary location on chip. The extracted filter may be used to predict transient temperature pattern for spatiotemporally varying power patterns by observing the digital programmable levels in Figure 51. The dynamic thermal variation is recovered from the digital on-off signatures. The recovered filter achieves 0.9846 correlation coefficient and exhibits good prediction of the sensor response given the digital input patterns. The implication of the measurement suggest the same technique may be used to monitor the system critical paths' response to transient operating condition by knowing the system workload through aggregating digital signatures such as block level power gating signal or DVFS states.

4.5.3 Fluidic Package Identification

The FPTE may determine the thermal field of an unknown package and die integration. A test structured is made for package level fluidic cooling solution. There has been many well know studies for thermal management for chips with relative large die size [97, 83, 62]. For die with sufficient surface area it is relatively straight forward to fabricate fluidic channel on the substrate for superior heat transfer as cited in Section 4.5.1. However, in a system-in-package (SiP) environment, the assembly becomes relatively difficult because the aggregated and signal routing on the same interposer with multiple carrying dies. The dies may have different thickness due to stacking (i.e. 3D DRAMs) and the thinner die will need to fill gap to the heat-spreader with thermal interface material (TIM), which may have relatively low in thermal conductivity on the heat extraction path. Further, the package level fluidic integration targets small die with high thermal density are not well understood.

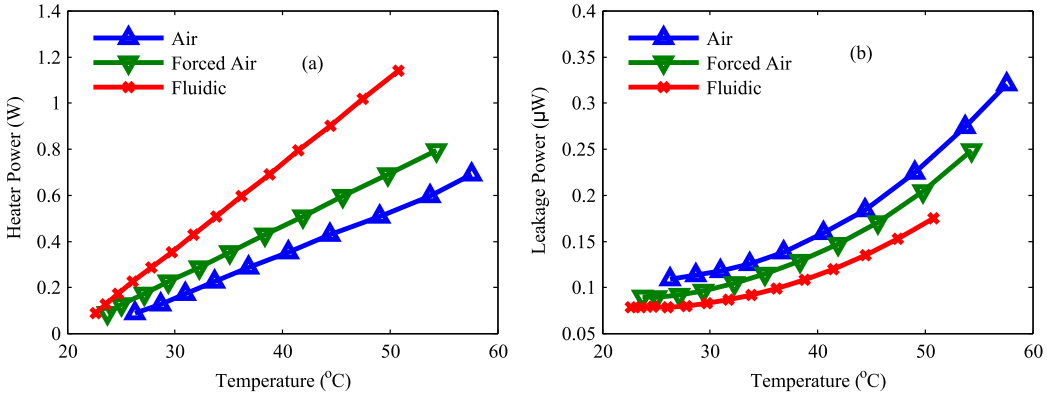


Figure 48: The measurement result of the FPPE system with configured power. (a) The exposed die with hotspot heating and the heat transfer. (b) the associated leakage power on die for each of the cooling methodology.

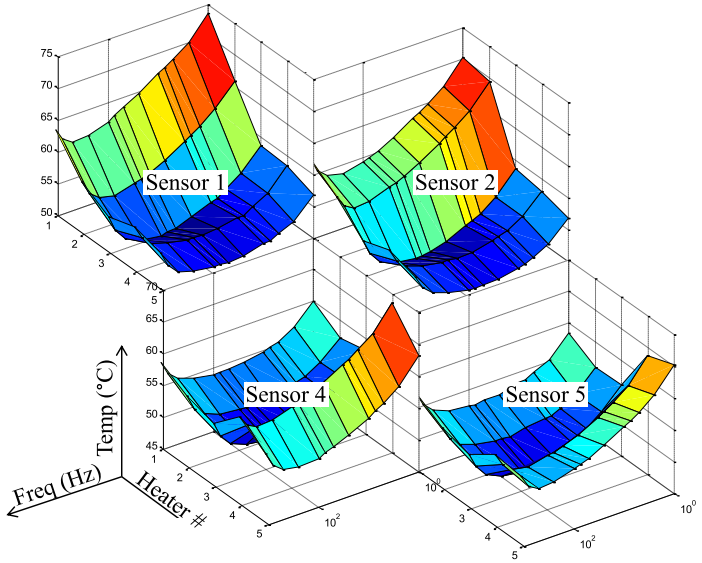


Figure 49: The filter response for given stimuli heater in magnitude.

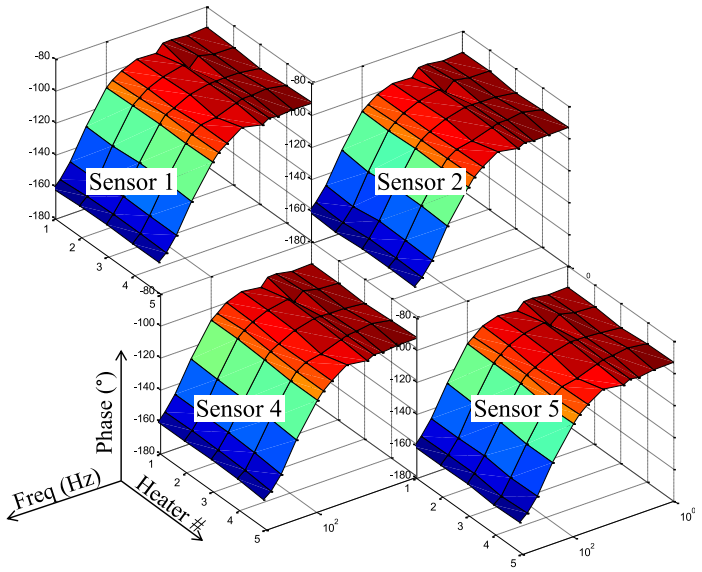


Figure 50: The filter response for given stimuli heater in phase.

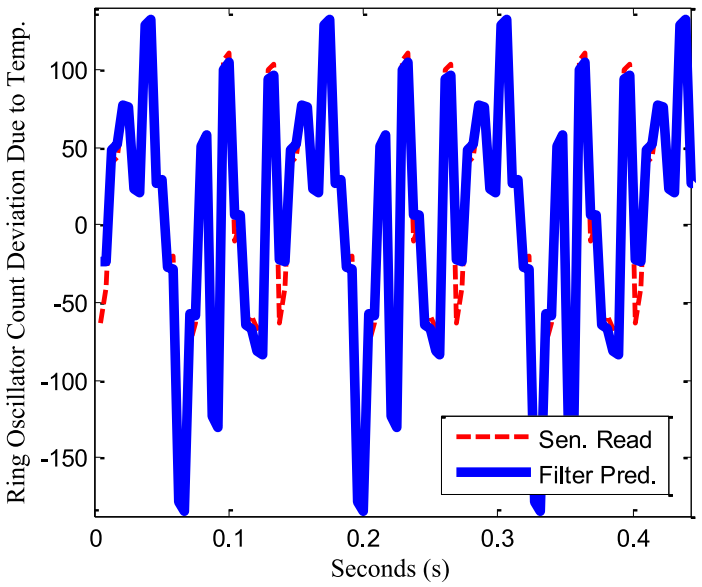


Figure 51: The measurement results showing time-varying arbitrary power pattern: (a) The applied power pattern and (b) variation in the sensor output.

The designs such as power regulator on board benefits from well designed on-board thermal management, and thermal vias were often mandatory for high power GaN FETs [96]. When the thermal management migrated in package, the die-to-die integration needs to consider the difficulty in channel formation and die handling. The alternative to fluidic channel – fluid-filled package integration – is generally a cost-effective solution for such integration without significantly penalizing the yield and integration cost. However, the fluid flow and non-uniform SiP surface may penalize the system modeling effort. The system is harder to model because of the different packaged die geometries and flow rates. To demonstrate the capability of the FPTE, a relatively simple fluidic integration may be formed inside the chip package in Figure 52. For the test chip, we applied cyanoacrylate coating to the open cavity and the bondwire. The cavity is then covered with acrylic cover with embedded fluidic inlet and outlet. The integrated cavity is sealed with the silicone sealant. The integrated system uses de-ionized (DI) water for the coolant.

The FPTE determines the system response under fluid filled chamber with constant fluid flow rate of 7 ml/min. The response of the sensor and heater pair are collected with the same methodology in [44]. The filter may be extracted with the same methodology as the system with traditional cavity. The fluid filled system is observed in Figure 53 for amplitude response and in Figure 54 for phase response. The amplitude response reduce from the system in Figure 49 shows the temperature coupling for the nearest heater and sensor pair reaches 75 degree temperature raise and 50 degrees raise at the uncorrelated heater and sensor pairs. The active fluid cooling reaches roughly 22 degrees in the strong coupled pairs and 13 degree at the uncorrelated pairs. The correlated pair thermal coupling improved by 3.4 times with simple fluid chamber configuration. The uncorrelated pair improves by 3.8 times and suggest the hotspot receive benefit from the fluidic heat spreading. The fluidic cooling improves heat spreading to the uncorrelated locations. Without modification

from the normal package configuration the extracted filter may be used to predict transient temperature pattern for spatiotemporally varying power patterns by observing the digital programmable levels in Figure 55. The recovered filter achieves 0.8739 correlation coefficient. The prediction accuracy drops comparing to the dry package scenario. The limitation may contribute to the temperature magnitude of the cooled system. Because the relative amplitude is roughly 1/10th of the dry package configuration. The system may be lesser accurate due to the quantization of the measurements. The evaluation is still satisfactory for coarse-gain tracking of device operating condition. The system may be used to determine a chip's thermal condition and coupling without prior knowledge of the packaging methodology and cooling method.

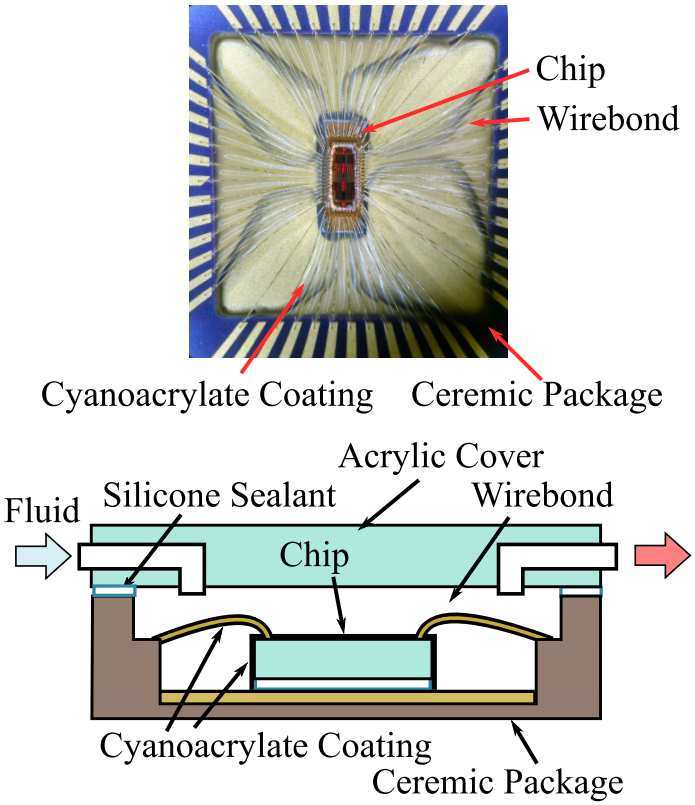


Figure 52: Schematic and experimental assembly of package, fluid chamber, and cover.

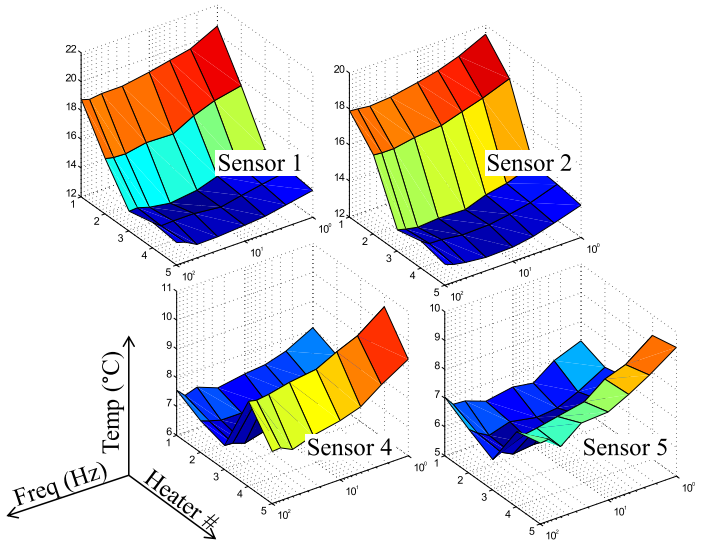


Figure 53: The filter response for given stimuli heater in magnitude.

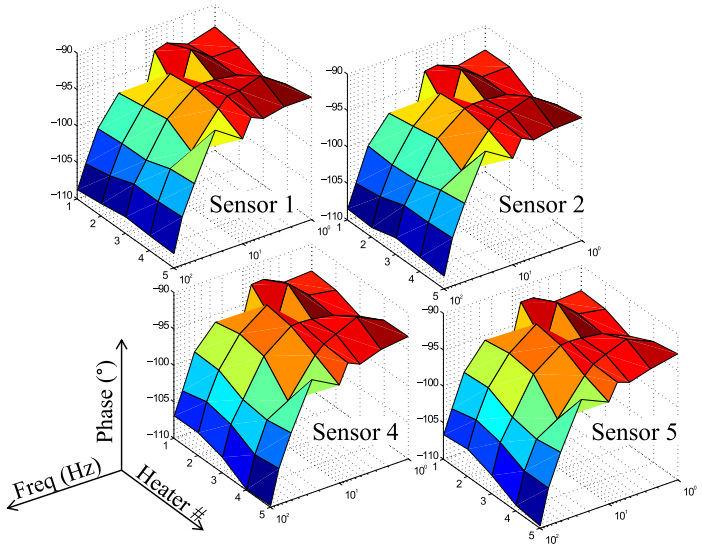


Figure 54: The filter response for given stimuli heater in phase.

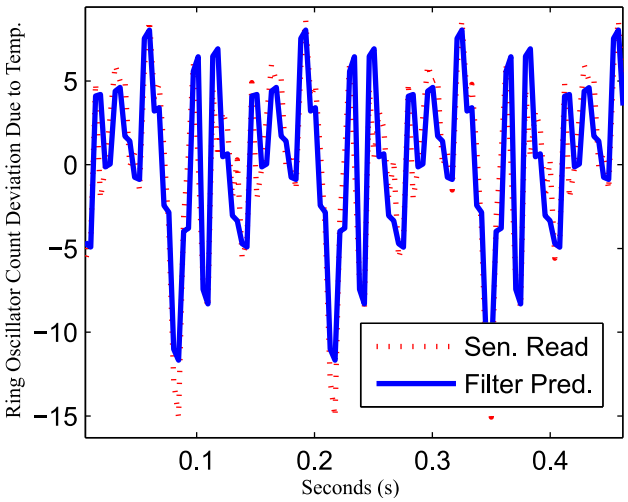


Figure 55: The measurement results showing time-varying arbitrary power pattern: (a) The applied power pattern and (b) variation in the sensor output.

4.5.4 Simulation Model Calibration

The FPTE emulation is effective for post silicon model correction. Due to metal stack density and the uncertainty in integration and calibration, it is often difficult to model system with sufficient confidence for practical system evaluation. To construct high confidence transient coupling model, the FPPE system may be used to calibrate a given high level model for a system integration model. To demonstrate the capability of this function, the thermal grid simulation in Chapter 2 is used for demonstration. The die and package material configurations in Table 1 was first used to provide system prediction in Figure 56(a). The die size for the model has been reduced to 1 mm by 2 mm and the die thickness has been increased to 200 μm to match the die dimensions. The periodic heating pattern in Section 4.5.2 was used to excite the modeled thermal grid. In simulation, the transient condition matched with the experiment poorly. The lateral thermal conductivity for the back-end-of-line (BELO) model is increased empirically and the overall thermal capacity is reduced to match the sensor readings in Figure 56(b). The difference between the original model and calibrated model is that the horizontal coupling is relatively stronger in the measurement than the early system. The uncertainty introduced from the metal density and the die-to-package attachment were corrected accordingly. The sensor #1 reading is shown along the simulation output in Figure 57. The result shows simulation framework can correctly predict the effect of the transient power to temperature variation.

4.6 Summary

This chapter presents the design and measurement of the field programmable thermal emulator (FPTE). The FPTE uses on-chip digital heaters and sensors to emulate time-varying power patterns on-chip, generate the corresponding spatiotemporally varying temperature pattern, and characterize the resulting variations in circuit

properties, for example, delay.

The FPTE has demonstrated the ability to generate emulated thermal patterns by programming heater registers accordingly. The effect of thermal coupling has been captured with the test setup. Having a programmable FPTE test-vehicle allows designers to understand the thermal effect on an architecture and/or workload considering electrical-thermal interactions, but without complete design/fabrication of the functional chip, thereby improve the design turnaround time.

As a thermal test-vehicle, FPTE has the advantage over existing thin-film heater approaches due to its compatibility with standard CMOS process, ability to generate controllable and time-varying power patterns, and directly characterize the effect of temperature patterns on device characteristics. The FPTE has been used to identify leakage and temperature correlation and transient thermal response for an air cooled system versus a fluidic cooled system. This experiment shows the potential for advanced package evaluation with actual leakage figure and device characteristics.

The FPTE has been demonstrated to be programmed on-line and characterize the thermal response of a packaged IC in experiments. This captures the exact thermal characteristics of a specific design-environment interaction considering process variations as well as time-dependent degradation in the thermal properties. Thermal filters were extracted through the FPTE structure for an air filled chip packaging cavity and a fluid filled chip packaging cavity. The recovered filter was used to reconstruct the thermal response of an arbitrarily generated power pattern at an desired location on chip. The correlation of the recovered thermal response and measured thermal response shows close resemblance. This capability allows thermal condition prediction through signature signals, such as the block-level enable signal or clock gating signal, at an arbitrary location on chip.

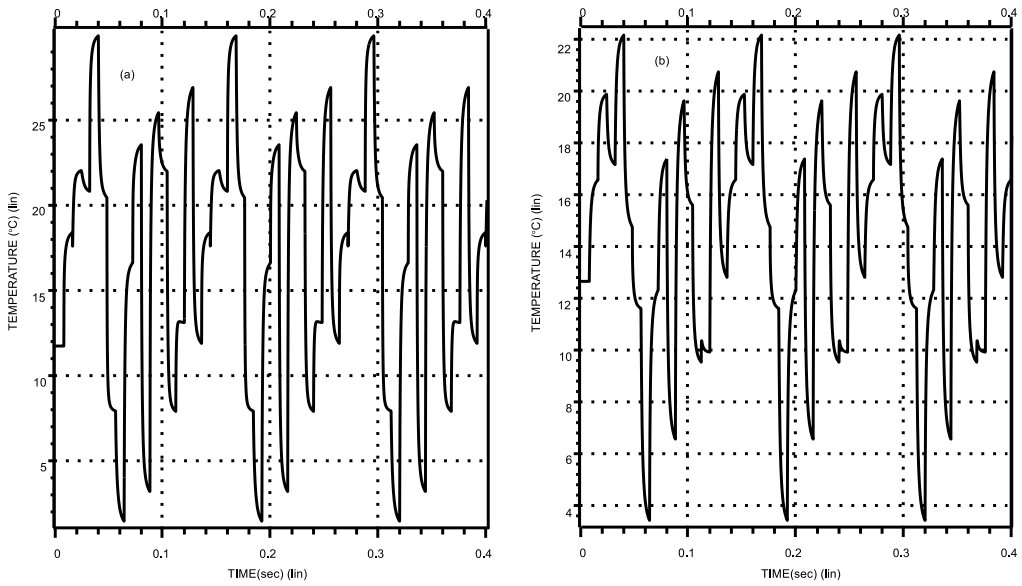


Figure 56: Package and die simulation without the transient calibration is shown in (a). The empirical fit on the BELO layer and die-attach in (b) shows relatively accurate modeling for the experimental FPTE die after calibration.

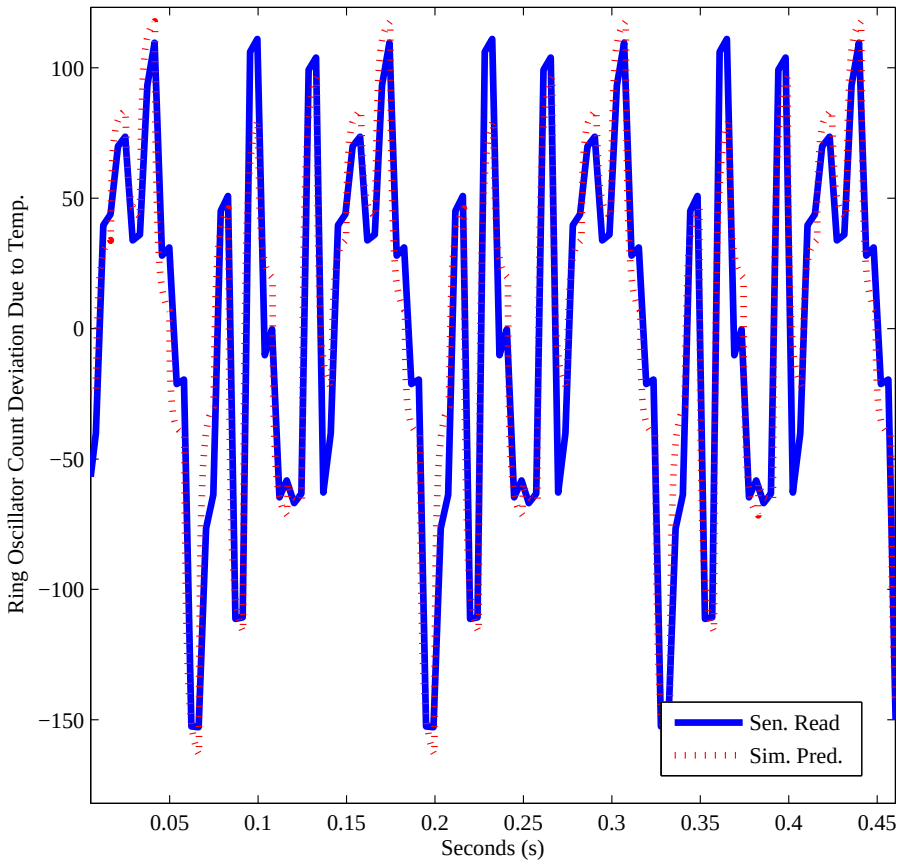


Figure 57: The superimposed figure between the simulation model and sensor reading.

CHAPTER V

CONTROLLING ELECTRICAL-THERMAL INTERACTIONS USING IN-PACKAGE MICRO-FLUIDICS

5.1 Introduction

In Chapter 2 and 3 we focus on electrical-thermal modeling and identification. In Chapter 4, we design a thermal test structure to supplement the conventional electrical-only BIST structure. In this chapter, we will apply thermal control structure to an embedded board to improve system power and bandwidth.

Maintaining the steady-state operation temperature has becoming one of the primary performance limitation to modern computation. For high performance computing, fluidic cooling and submerged cooling have becoming increasingly popular in server farms and data centers [21]. The demand for computation with more functionality at elevated efficiency has led the industry to adopt the active fluidic solutions [58]. However, on the mobile segment, due to the integration challenges, the thermal management aspect still lags behind the state of the art active fluidic solutions in the other system segments. The computation capabilities of the system-on-chips (SoCs) have advanced significantly in recent years. The embedded SoCs are expanding their applications beyond smart-phones and tablets, for example, in robotics, autonomous avionics, and internet of things (IoT) [88, 24]. With increasing computation demand and processing capabilities of these systems, the thermal management may emerge as an unexpected challenge for common use cases [35]. The unmanaged high temperature reduces performance, increases device leakage, and accelerates device aging [75, 10]. The requirements on form-factor, low power, and user experience may lead to a more

aggressive cooling technology for SoCs.

First, we present an in-package microfluidic cooling technology using micro pin fins. The structure is embedded in a chip-scale silicon interposer that can be separately fabricated and integrated with the die during packaging, without interfering with the chip fabrication. The direct die-attachment is compact and improves the heat extraction capacity compared to external cooling. The in-package cooling technology is fabricated in-house and attached to a commercial SoC (Snapdragon 600, Figure 58). The details of the fabrication process of the interposer with integrated fluidic pin-fins can be found in PhD thesis of Z. Wan, Georgia Tech, and in the article [86]. A low-power piezoelectric pump, controlled by the SoC, is integrated with the system. The experiments are performed considering single-phase cooling with deionized (DI) water. The system level temperature, power (processor and pump), and performance are measured considering benchmark applications running on the SoC. The results are compared against the external (on-package) passive/active heat removal technologies.

The experiment demonstrates that the in-package fluidic cooling system can reduce the SoC energy consumption and integration footprint. The measurements with benchmark applications showed that the in-package cooling operated at ˜30 °C lower temperature, ˜24 % lower energy, and ˜30 % better performance compared to the baseline (no cooling) SoC. Compared to the external passive cooling, in-package cooling reduced peak temperature by ˜20 °C and peak energy by ˜16 %. The energy analysis considers the pump power (peak 110 mW). The SoC with in-package cooling has 2.5 X and 3 X lower footprint compared to the passive cooling and external fluidic cooling, respectively.

5.2 Related Work

Active cooling with microfluidics has been studied for high performance computing [83, 62, 5, 28, 47]. The state-of-the-art methodology is capable of integrating fluidic systems on a silicon substrate or an interposer [47]. Applying the fluidic heat-sink on the chip package improves heat-removal capacity. Forming fluidic channels and micro-pin fins directly on the die may leads to even more effective cooling [98]. Tuckerman et al. and Peles et al. reported impressively low thermal resistance measurements of 0.09 K/W and 0.0389 K/W, respectively [83, 62]. However, fabricating channels/pin fins directly on the chip requires foundry support, which is difficult to adopt for relatively cost-sensitive systems. On the other hand, the less effective external "on-package" heat-sinks increase the system board's keep-out area for the SoC. The ease of integration, the system footprint, and the power dissipation associated with active cooling are additional criteria to determine a suitable cooling solution. This chapter experimentally demonstrates an in-package fluidic cooling for a commercial SoC in Figure 58. The prior work on in-package fluidic cooling had concentrated on experiments with dummy silicon dies/heaters [83, 62, 5, 28, 47]. The simulation based evaluation has shown the fluidic cooling improves the system level energy-efficiency [76, 69]. However, to the best of our knowledge, experimental characterization of in-package fluidic cooling on a fully functional commercial SoC has not been reported in the literature. Moreover, prior work mainly focused on the high power-density systems, the role of fluidic cooling in embedded SoC has not been investigated.

Simulation based evaluation of system level energy-efficiency advantages of fluidic cooling has been reported. For example, Sridhar et al. have considered fluidic cooling microarchitecture simulation of the 3D ICs [76]. The work by Serafy et al. has speculated that energy reduction may be achieved through chip stack cooling simulation considering the pump energy [69]. He et al. further modeled a high performance

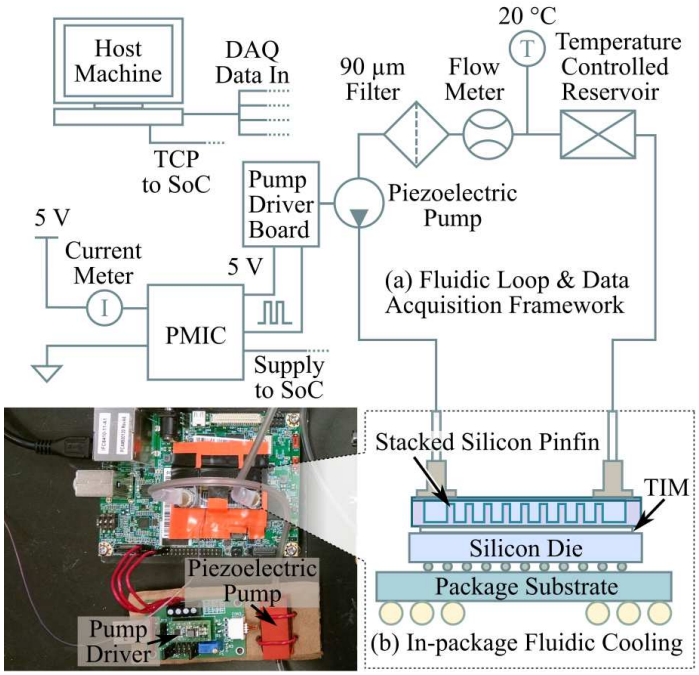


Figure 58: The experimental characterization of the in-package fluidic cooling: (a) A schematic of the measurement setup. (b) A snapshot of the board and pump assembly used for the measurement. The in-package fluidic cooling is integrated with the SoC (Snapdragon 600). The piezoelectric pump's driver circuit directly draws current from the PMIC on the IFC6410 board. The hall-effect sensor measures the total current entering the board. The driver circuit takes PFM frequency control signal from the programmed GPIO pin.

system with a more advanced pin fin cooling channel [92].

Because the physical properties of the air convection is fixed with given form factor, majority of the designs focus on heat spreading within the housing enclosure and improve cumulative specific heat. There have been many studies on advanced material on small form-factor cooling stack [85, 49]. A known recent high performance mobile device has already integrated heat pipe for thermal management [84]. In order to sustain the performance and maintain contact temperature, nano-materials such as graphite sheets were utilized to improve heat diffusion over device surfaces. Phase change material for improve sprinting duration is also reported in few literature. Other techniques such as shaping the electricalmagnetic interference (EMI) shield on die to form cavity pocket avoiding direct path from hotspot to enclosure surface has been reported by [70]. The thermal gradient across the surface is difficult to improve with passive heat spreading by nature. High thermal conductivity material may also be infused in the coolant to enhance the convection heat transfer [26]. Measurements independently reported by Gurrum et al. and Wagner et al. reported the average tablet surface temperatures are roughly 35 °C when the enclosure hotspots are at 41 °C or above [85, 27].

5.3 System Integration

5.3.1 Embedded SoC Platform

The platform for embedded SoC evaluation was the IFC6410 board from Inforce Computing. The embedded 28 nm SoC on the board is a Snapdragon 600 from Qualcomm with an Adreno graphics engine and an up-to-1.9 GHz quad-core system. The flip-chip bonded SoC has the die's backside exposed, which allows direct silicon fluidic attachment. The system runs on the operating system Linaro-Gnome, a Linux ARM distribution. The SoC's built-in temperature sensors collect thermal information during operation. Out of the 13 thermal sensors on-board, only four sensors for

processors are considered for the analysis. The thermal information is averaged into one aggregated reading and transmitted to a desktop server at a fixed-one-second interval. A TCP communication client updates the value to the connecting server on the network. Along with the thermal information, the core clock frequencies are also uploaded with the same framework. The reporting framework roughly consumes 2 % core one's bandwidth in the background. The reporting mechanism along with various background processes increase core one's temperature by 2 °C higher than the remaining three cores.

5.3.2 Integration of Cooling Technology with SoC

The SoC with integrated in-package silicon-based active fluidic cooling and the baseline systems has been constructed. The in-package fluidic system is benchmarked against the bare die, natural convection heat sink, and conventional fluidic cooling. No Cooling: The exposed die configuration came with the original system. The SoC relies on thermal throttling and frequency scaling to achieve thermal management. The configuration reaches 70 °C under a moderate workload without external intervention. The system configuration is shown in Figure 59(a). Proposed Active In-Package Fluidic Cooling: The micro pin-fin system is integrated with the Omegatherm 201 thermal paste between die and heat sink. The system was secured with electrical tape on top of the die. The micro-fluidic-pin-fin region aligned with the active SoC die area. Note in Figure 59(b) that the tubing and 20 mm extended connector areas on each side are not necessary for a commercially integrated in-package fluidic cooler. The connection would be routed through the fluidic trace embedded in the printed circuit board (PCB) [47]. Baseline Passive Air Cooling Heat Sink: The natural air convection relies on the heat sink surface area to remove heat from the SoC. The attached aluminum heat sink has a dimension of 40 mm 50 mm with fin height 30 mm. The heat sink was designed for the AMD RS780L chipset with 10 W peak power.

The assembly is shown in Figure 59(c). Baseline Active External Fluidic Cooling: An external copper fluidic cooler with a dimension of 50 50 mm is considered. The cooler contain 4 mm 4 mm cone-shaped pin fins and the channel height of 10 mm. The assembly is shown in Figure 59(d). The cold-plate contact surface area is larger than the chip area and increases horizontal footprint and vertical height of the SoC. The copper head is hence mounted tilted to make leveled contact with the die. While there are no reported embedded systems that apply full sized fluidic cooling, we consider this configuration as a baseline external fluidic cooling system.

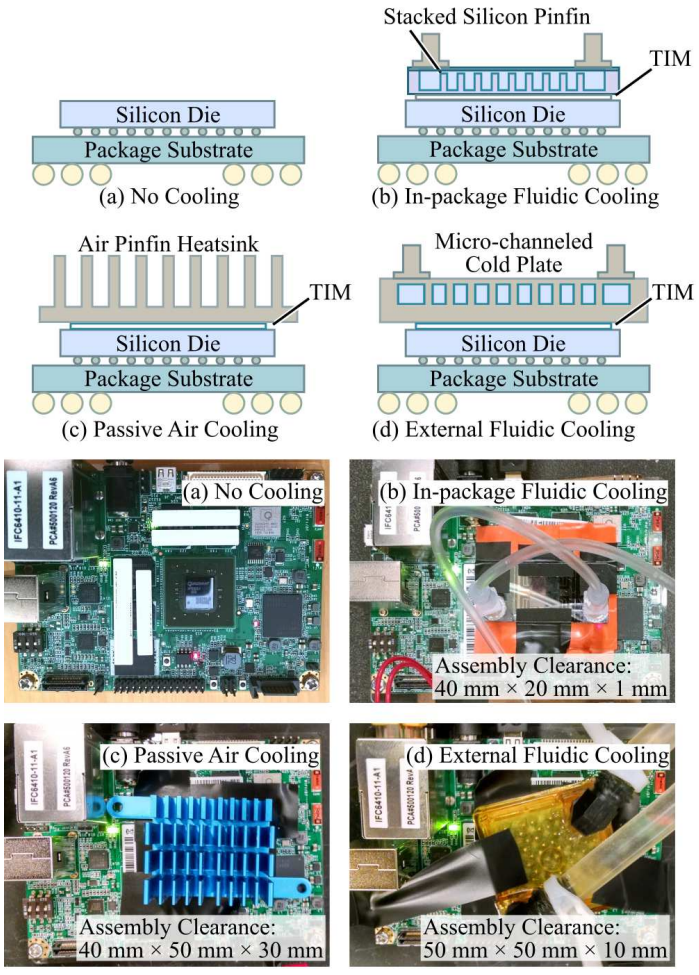
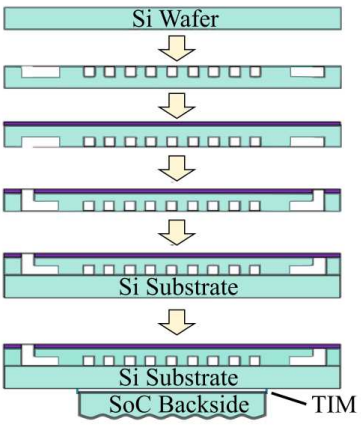


Figure 59: The schematic and pictures of different cooling options: (a) an IFC6410 board without cooling, (b) the proposed in-package fluidic prototype mounted on the die, (c) the same board with passive air cooling solution, and (d) the external fluidic cooling solution.

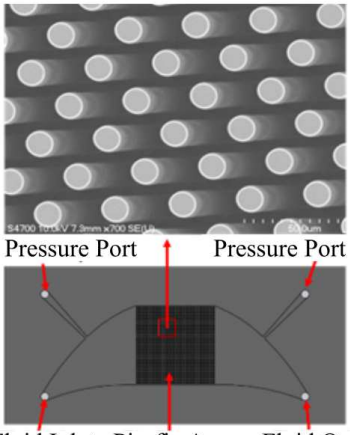
5.3.3 In-package Fluidic Cooling

Incorporating a heat-exchange layer in a direct contact to the SoC reduces the overall thermal resistance. In conventional external (on-package) cooling, the thermal solution is integrated during component assembly. Traditional metal heat sink may not directly attached to a silicon die due to potential electromagnetic interference (EMI) challenges and the mismatch in the thermal expansion coefficients. We present a silicon based fluidic interposer designed to be permanently attached to the die and enclosed inside the EMI shield in Figure 60(a). *The concept, design, and fabrication of the micro pin-fin array is contributed by by Z. Wan and Y. Joshi, from School of ME, Georgia Tech.* The micro pin-fin array was fabricated in the clean room and is a contribution from [86]. *The concept, design, and fabrication of the micro pin-fin array is contributed by by Z. Wan and Y. Joshi, from School of ME, Georgia Tech.* The microgap fabrication process started with a double-sided polished 4" silicon wafer with a thickness 500 μm. In the first step, positive photoresist SPR-220 was spun and exposed to form a mask of the microgap. Then the wafer was etched in the deep reactive ion etching (DRIE) process. Using the standard Bosch process, which alternates between a plasma etching step and passivation step, the deep microgap cavity with staggered micro-pin-fin array was etched. Tencor P15 profilometer was used to record the depth of the microgap. In the second step, the wafer was flipped and a 2 μm thickness silicon oxide layer was deposited by plasma enhanced chemical vapor deposition (PECVD) method as an insulation layer.

In the third step, the wafer was taken through a photolithography step and a reactive ion etching (RIE) process to remove the oxide and expose the silicon which was to be etched to form the fluid vias. After the RIE process, the wafer was put into DRIE to continue to etch the silicon and developed the fluid vias. Thereafter, the processed wafer was diced and the microgap samples were taken out of the wafer. In the last step, a 500 μm thick silicon wafer was bonded to the diced microgap samples



(a) Micro pinfin fabrication steps



Property	Dimension
Silicon Area	43 mm × 20 mm
Pin-fin Array	10 mm × 10 mm
Microgap	176 µm
Pin Diameter	17 µm
Space (Longitude)	45 µm
Space (Transverse)	40 µm

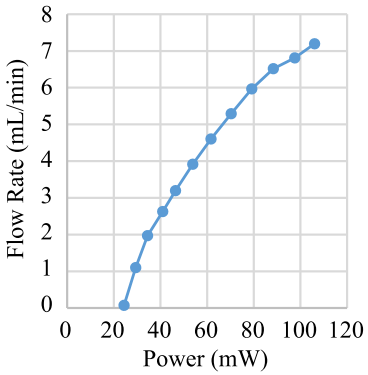
(b) Device SEM image and key paramters

Figure 60: The fabricated device and its corresponding features are highlighted in this figure. The key steps for micro pinfin fabrication are listed in (a). The parameters and the SEM image are shown in (b).

by epoxy to form a sealed device. The dimensions of the fabricated device are 43 mm 20 mm. The microgap also includes pressure ports at the fluid inlet and outlet, which are not used in the present study. The area of the pin fin array is 1 cm 1 cm. The depth of the microgap is 176 μm, the diameter of the pins is 17 μm, longitudinal spacing is 45 μm and transversal spacing is 40 μm. A fabrication flow is shown in Figure 60(a). The Figure 60(b) shows an image of the fabricated device with the SEM image of the staggered pin-fin array and the highlighted key parameters.

5.3.4 Piezoelectric Pump

The choice of pump for active cooling in SoCs is limited by the motor's physical geometry and power consumption. We considered compact high-flow-rate piezoelectric pumps for this investigation. As an example, the pump model MP6 manufactured by Bartels Mikrotechnik is chosen for this study. The pump dimension is 30 mm 15 mm 3.8 mm and the driver board dimension is 10.5 mm 20.5 mm 6 mm. The pump's peak power is 110 mW. The piezoelectric pump has peak pumping flow rate at 7 mL/min. The pump standby power is measured to be 10 mW. The piezoelectric pump's flow rate may be programmed with digitally controlled pulse frequency modulation (PFM) driver. While the pump itself has a wider operating range, the operation point of interest for this work is tabulated in Figure 61.



Property	Value
Pump Size	30 × 15 × 3.8 mm^3
Driver Size	10.5×20.5×6 mm^3
Power Supply	5 VDC
Frequency	15 – 100 Hz
Flow Rate	0 – 7 mL/min

Figure 61: The power characteristic and the corresponding parameters associated with the piezoelectric pump.

5.3.5 Integrated Fluidic Loop

The configuration of the fluidic loop consists of the pump, controlled temperature reservoir, flow meter, filter, and the SoC chip with the in-package fluidic cooling, see Figure 58(a). A swappable Cole-Parmer digital gear pump which is capable of flow rates from 5.52 mL/min to 331.2 mL/min was also used during calibration besides the MP6 pump. A part of the flowloop was immersed into a controlled temperature bath. A McMillan S-114 flow meter was calibrated to measure the volumetric flow rate. A 90-μm Swagelok filter was used to keep the inlet water clean and prevent clogging of the microgap and the piezoelectric pump. The PFM controllable micro pump was powered through the 5 V rail from the power management-integrated circuit (PMIC) on the IFC6410 board. The on-board GPIO controlled the pump's PFM clock source. The total power of the system including the board (SoC + peripheral) and pump was measured with TCP202 hall effect sensor on the 5V line near the board's power socket.

5.4 Experimental Observation on Fluidic Cooled SoC

A subset of Splash-2 benchmark suite was used to demonstrate the thermal behavior of the SoC with different cooling technologies considering the workload [91]. The Linux workload generator tool, Stress, was used to bring the thermal condition to the steady state [89]. In all benchmarks, the Stress tool spun sqrt() on all four cores. The benchmark also ran with four-process parallelism.

5.4.1 Pump Power And SoC Power Tradeoff

The leakage power at a higher die temperature may exceed the active cooling power. This suggests an active cooling system may consume lower power compared to a constrained passive cooling. The preceding hypothesis is validated in Figure 62. We first ran the Stress tool in all the four cores over a period of time that was long enough to reach a steady-state temperature. Next, we terminated the application

to reduce power and allowed the temperature to cool to a steady-state. The power dissipation of the board and the temperature of the SoC were measured by enabling and disabling the fluidic loop. A measured flow-rate of 7 mL/min was considered in the experiment. The sub-figures in Figure 62(a) shows the power and temperature when Stress was running. We observed the SoC heats up to 55 °C within one minutes of operation and 70 °C for 10 minutes continuous full load. When the fluidic loop was disabled, the system consumes additional 174 mW of power at the one-minute mark and additional 490 mW of power at the ten minutes mark. We believe the additional power was due to the increased temperature induced leakage. The Figure 62(b) shows the power and temperature during the low workload (idle) condition. With the fluidic loop 'turned-off', although the SoC temperature remains higher, the system power becomes lower than the case when pump was on. We believe this was because, the aggressive circuit/micro-architecture level idle power management techniques in commercial mobile SoCs significantly reduce the leakage current during low-workload conditions. Consequently, the overhead associated with the pumping power made the system power larger with the fluidic loop on.

5.4.2 Steady-State Temperature at Full Utilization

We first consider the full utilization scenario. The comparisons of system power, temperature, and footprint/height of the cooling solution are shown in Table 9. All measurements were performed considering the same Stress' workload. The fluidic channels were driven from the same-pump at fixed 7 mL/min for both external and in-package cooling. The system power includes the pump power. The measurement results shows that the in-package cooling can reduce temperature by 24 °C and hence, the leakage current, resulting in 450 mW power saving over passive cooling during the peak workload condition, even after accounting for the pumping power. The reduced temperature may have an additional benefit of slowing down the aging process of the

devices and may improve the lifetime of the SoC.

Table 9: The system steady-state power and temperature comparisons with assembly clearance

Cooling Structure	Symbol	Peak Temp.	Peak Power	Footprint	Height
No Cooling	nc	84 °C	6.28 W	N.A.	N.A.
Passive Air Cooling	ac	64 °C	5.83 W	2000 mm^2	30 mm
External Fluidic Cooling	fc	42 °C	5.53 W	2500 mm^2	10 mm
In-package Fluidic Cooling	ifc	40 °C	5.39 W	800 mm^2	1 mm

5.4.3 Application Dependent Power

The active fluidic cooling improves performance through avoiding overheating beyond the thermal design point (TDP). The transient temperature measurements were performed considering the 'Splash-2' benchmarks. The benchmark, Raytrace, which is known to be unfriendly to DVFS, was applied [39]. The Figure 63 shows the Raytrace application on the bare system and the system with the in-package fluidic cooling. The system without fluidic cooling had limited thermal headroom. Once the hot cores exhausted their sprinting budget in a short period, the internal power controller of the SoC forced the system to operate at a reduced power state to stay within the thermal threshold. Moreover, the thermal throttling also incurred a delay penalty. When active in-package cooling is used, we observed that the SoC stayed at a higher power state during the entire operation, without violating thermal constraints. Consequently, the application run time and the computation energy were reduced. Overall, we observed the system without cooling consumed 31.5 % more energy than the in-package fluidic cooling for the same workload. The chip peak temperature was also 32 °C higher.

Four systems were compared considering the benchmark FMM, which has the highest floating-point operations per second (FLOPS). The bare system without the additional thermal management was, again, thermally throttled to a lower power state during the execution and took more time to complete. Hence, it consumed

more energy in Figure 64. All the systems with cooling (passive, external fluidics, and in-package fluidic) sustained the maximum operating power without throttling. However, the temperature with external passive cooling was higher than the other two cases. The power and temperature characteristics of the external fluidic cooling are similar to the in-package cooling, but as noted before, the external fluidic cooling takes 3 times larger footprint and 10 times taller in height, than the in-package cooling with the micro pin-fin. The measurements showed that compared to the in-package fluidic cooling, the bare system consumed 28.9 % more energy, while the system with passive cooling consumed 18.3 % more energy during execution. The peak power for the system with passive cooling was also 540 mW higher than the in-package fluidic cooling case mainly due to higher leakage at an elevated temperature. The measurement results of the temperature, completion time, and energy dissipation for various Splash 2 benchmarks are summarized in Figure 65. The energy dissipation was calculated considering the total power (board + pump) and the completion time. All the benchmarks listed in the figure were below one minute of run time. The cumulative leakage power was more apparent in longer benchmarks. In all the benchmarks, the energy advantage of the in-package fluidic cooling over the no cooling and passive cooling systems was apparent. When compared against the external fluidic cooling, the in-package cooling shows similar energy for shorter benchmarks and occupies 32 % less footprint. In a throughput based benchmark, like Raytrace, or benchmarks with significant higher FLOPS (FMM and Barnes) the in-package fluidic cooling showed energy advantages over the external fluidic cooling. Further, the DVFS friendly benchmark Ocean_cp (because of its process synchronization barriers) still consumed 13.7 % more power in the passive cooling compared to the in-package fluidic cooling.

5.5 Close-loop Thermal Management using In-package Fluidic Cooling

5.5.1 Surface Hotspot Mitigation

Because the physical properties of the air convection is fixed with given form factor, majority of the designs focus on heat spreading within the housing enclosure and improve cumulative specific heat. There have been many studies on advanced material on small form-factor cooling stack [85, 49]. A known recent high performance mobile device has already integrated heat pipe for thermal management [84]. In order to sustain the performance and maintain contact temperature, nano-materials such as graphite sheets were utilized to improve heat diffusion over device surfaces. Phase change material for improve sprinting duration is also reported in few literature. Other techniques such as shaping the electrical-magnetic interference (EMI) shield on die to form cavity pocket avoiding direct path from hotspot to enclosure surface has been reported by [70]. The thermal gradient across the surface is difficult to improve with passive heat spreading by nature. High thermal conductivity material may also be infused in the coolant to enhance the convection heat transfer [26]. Measurements independently reported by Gurrum et al. and Wagner et al. reported the average tablet surface temperatures are roughly 35 °C when the enclosure hotspots are at 41 °C or above [85, 27].

5.5.2 Enclosure heat sink Design

Limited by the system-mobility criteria, the heat spreader to ambient area is confined to lesser or equal to the housing enclosure. The on-die fluidic cold plate may serve as both sprint computing heat buffer and hot-spot spreading layer. The active cold plate on the SoC should be able to carry heat away from the chip and to shield SoC heat from propagating through the enclosure surface unevenly.

In a hand-held system, the SoC thermal design point (T_{SOC}) has a has an upper

limit of 90 °C. The peak SoC power (P_{SOC}) is roughly at 8 W. The enclosure surface temperature (T_S) is limited by touch temperature at 41 °C and is a function of the vertical cumulative thermal resistivity and lateral diffusive thermal resistivity. In order to satisfy enclosure contact temperature and the SoC's TDP, vertical cumulative resistivity is bounded by

$$VerticalR_{TH} < \frac{T_{SOC} - T_S}{P_{SOC}} \tag{4}$$

.

where TSoC <90 °C, TS <41 °C, and TSoC <8 W. The constraint limits the vertical R_{TH} <6.1 K/W, which is not difficult to achieve even for mobile form factor. Low cost housing material such as acrylic glass or plastic may still be used at millimeter thicknesses. The more stringent limitation for hand-held systems is the lateral diffusive resistivity from the SoC to the edge of the enclosure. The limitation may be significantly improved by forced convection. In this work, an acrylic based cold plate and an aluminum based cold plate are designed 66. The plate area is 60 mm x 132 mm and the fluid area is 50 mm x 111 mm x 0.5 mm. The pin-fins are 2 mm x 2 mm squares. The longitudinal spacing is 4 mm and transversal spacing is 3 mm. The fins are designed for structural support. The acrylic based design demonstrates the possibility of the display side cooling. The aluminum based design resemble the back side cooling. Both plates are covered with plain 1 mm acrylic sheet. The measured temperature difference for aluminum plate cooler between the inlet and outlet are 30.6 °C and 28.3 °C respectively at the peak 4.8 W board power. The acrylic based fluidic sink is 31.3 °C and 29.6 °C. The temperature percentage difference between metal and acrylic material is less than 3 %. This demonstrated the passive heat exchange for low power system dominated by air and the high thermal conductive material is not of a major concern. Further, the forced convection shows lesser than

8 % surface temperature gradient across the inlet and outlet for the aluminum cold plate and 6 % for the acrylic cold plate.

The same benchmarks for Subsection 5.4.3 is applied in this experiment. The baseline case with no cooling, the passive heat sink cooling, and the microfluidic cooling are introduced. On the radiator side, the constant temperature bath (in-package fluidic cooling – ifc), the acrylic cold plate (in-package fluidic cooling closed-loop – ifcc), and the Aluminum cold plate (in-package fluidic cooling metal-heat sink closed-loop – ifcmc) are benchmarked. The time to completion is shown in 67. The computation energy is shown in 68. The temperature is shown in 69. For longer benchmark, the ideal heat exchanger appears to outperform the acrylic and metal cold plate. For all the benchmark, the improvement over passive cooling is observed in all three fluidic cooling configurations. The temperature improvement is significant for fluidic cooling and with marginal difference between cold plate of choice. The passive cooling energy is still signification during active core operation and the total power exceed active pump energy considering the cold plate efficiency.

5.5.3 Cooling Threshold Management

The fluidic cooling is effective methodology of bringing heat on the enclosure surface and reducing in-package temperature. The active cooling system takes advantage of the on-chip power reduction to power the pump and still achieving lower system level power. Further control analysis has been performed to balance the pump power and on-chip temperature. Few simplistic control mechanisms are benchmarked: always on (ao), maximum temperature threshold (mt), and maximum frequency threshold (mt).

Always on: The always on control scheme is equivalent to the in-package fluidic cooling experiments where the heater constantly driven to the maximum frequency and flow rate.

Maximum temperature threshold: A bang-bang controller is implemented for the fluidic pump in software. The micro-processor completely shut-off the pump's pulse frequency modulator signal and power gate the pump's boost regulator for minimal sleep power consumption. The temperature threshold for each processor is polled every second and when any processor's temperature exceed the temperature threshold, the enable signal turns on the pump and the boost-regulator. The temperature threshold is selected to be 55 degree Celsius.

Maximum frequency: A similar bang-bang controller is implemented for the fluidic pump in software. The frequency threshold for each processor is polled every seconds and when any processor's frequency exceed the frequency threshold, the enable signal turns on the pump and the boost-regulator. And frequency threshold is selected to be 1 GHz.

The steady state power consumption shows less than one percent of difference between each cooling methodology at full load.The time to completion is shown in 70. The computation energy is shown in 71. The temperature is shown in 72. The energy benchmark shows the energy saving applying different technique is not significant. The major reason is the temperature in all the cooling objective are not producing significant temperature difference. The pump power itself also is relative trivial comparing to the board power, consuming roughly 2 percent of the total power. Any improvement in the pump power will be less prominent during active computing. During quiescent state the frequency based activation achieve 79 mW of power reduction because pump is mostly offline. The temperature based activation achieve 41 mW of power reduction due to constant temperature upkeep. Comparing to the overall power dissipation, the techniques reduce 1.5 to 2.8 percent of the overall power under DVFS. As an energy conscious cooling policy, the frequency based cooling should be employed because the benefit of lower quiescent power and takes advantage of the DVFS leakage reduction. Theoretically the frequency based cooling

will be penalized by the latent heat of the previously runs and have worst performance due to throttling, however because the in-package fluidic system has relative high flow-rate considering the cavity's cross-section. The response time of the control algorithm allows comparable system performance to the temperature based control. The temperature threshold based control does not seems necessary for fluid driving control especially the fail safe thermal throttling is designed through DVFS frequency control. Without co-designing the frequency scaling opportunities, it appears simply following the frequency threshold for pump driving improves energy trade-off.

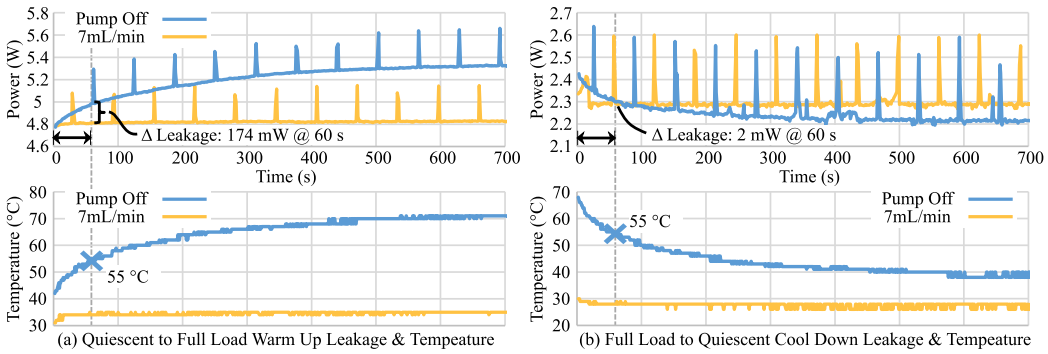


Figure 62: The measurements show the system power and the SoC temperature following enabling/disabling of the fluidic loop in the in-package cooling technology. (a) At a high workload condition with full utilization of the cores, the system without active cooling operates at a higher temperature and sustains a higher leakage. The active cooling reduces temperature, and hence, leakage, to reduce the total system power even after accounting for the pumping power. (b) On the other hand, in the idle or low utilization condition, the SoC employs aggressive idle power management to electrically minimize leakage power; consequently, the temperature reduction with the active cooling does not translate to power saving. The pumping power overhead makes the fluidic cooling less efficient. The measurement shows the need to couple electrical power management techniques with active fluidic cooling for an optimal power management system targeting low power SoCs.

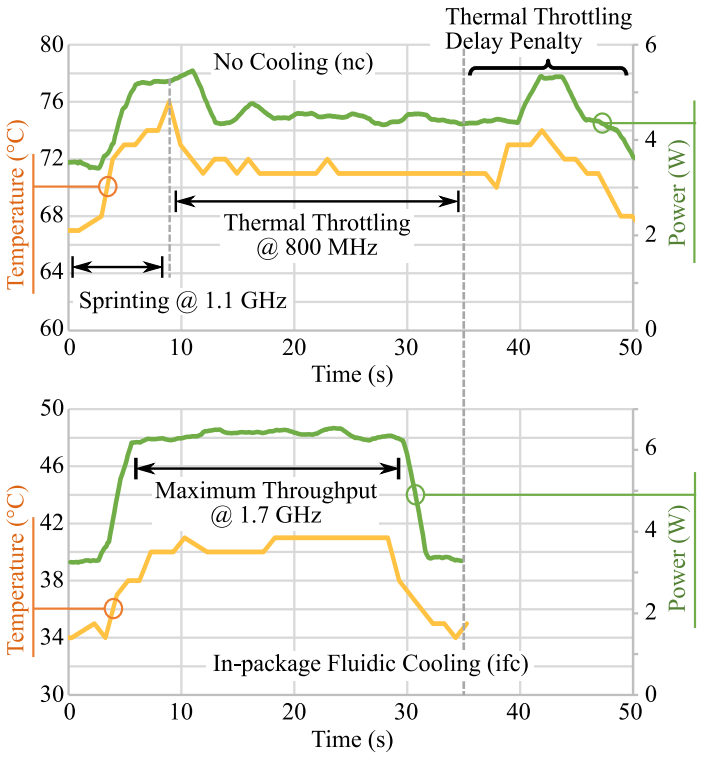


Figure 63: The measurement results of the temperature and power characteristics with the bare-die (no-cooling) case and the in-package-active-cooling case. The traces were collected from the benchmark "Raytrace." Without any thermal management, the higher temperature limited operating time in high performance (high-power) mode and induced throttling, thereby increased the computation time. The higher computation time led to higher energy dissipation. The system with the active in-package cooling ran at a higher power mode without throttling resulting lower computation time and, hence, lesser energy dissipation.

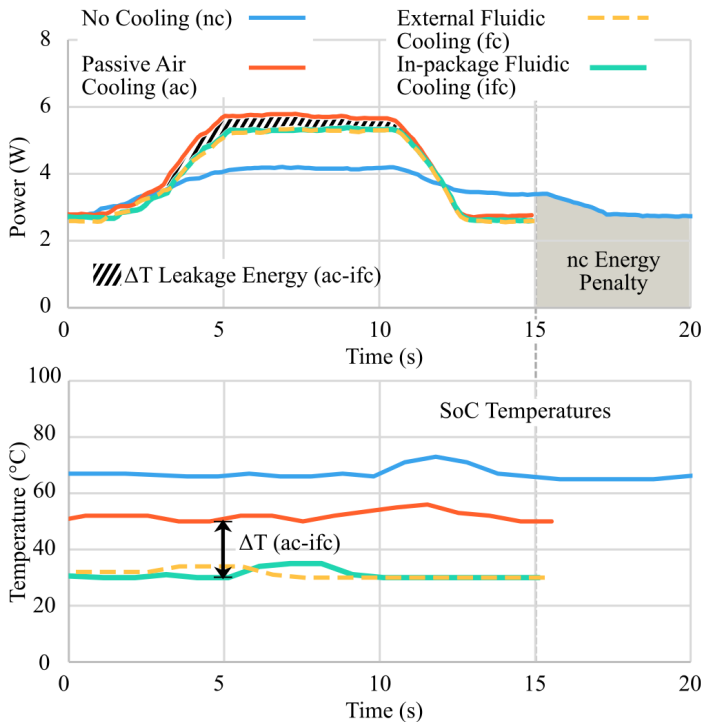


Figure 64: The measurement results of the power and temperature responses with different systems running the benchmark "FMM." The bare system with no cooling was forced to operate at a lower power/performance mode due to a higher temperature and a higher completion time. The passive air-cooled heat sink prevented the thermal throttling but a higher temperature lead to a higher power (higher leakage). The in-package and external fluidic cooling showed similar performances but the in-package cooling had a much smaller footprint/volume.

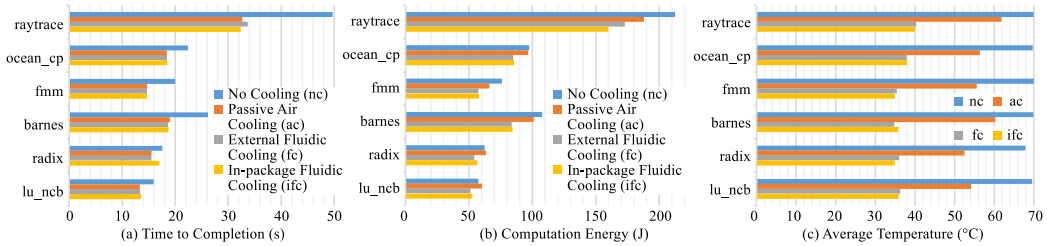


Figure 65: The measurement results for various Splash-2 benchmarks running on the SoC for (a) the completion time, (b) the total computation energy, and (c) the average temperature.

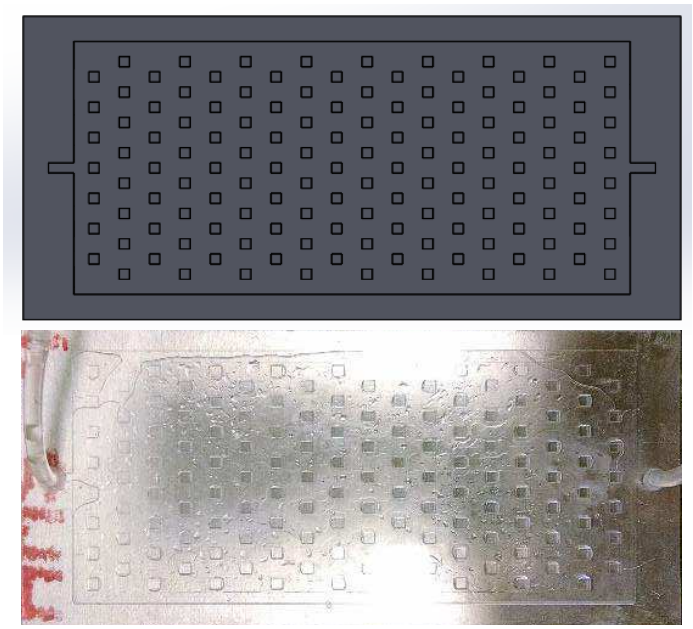


Figure 66: The fluid to ambient cold plate's mechanical drawing and machined assembly is shown.

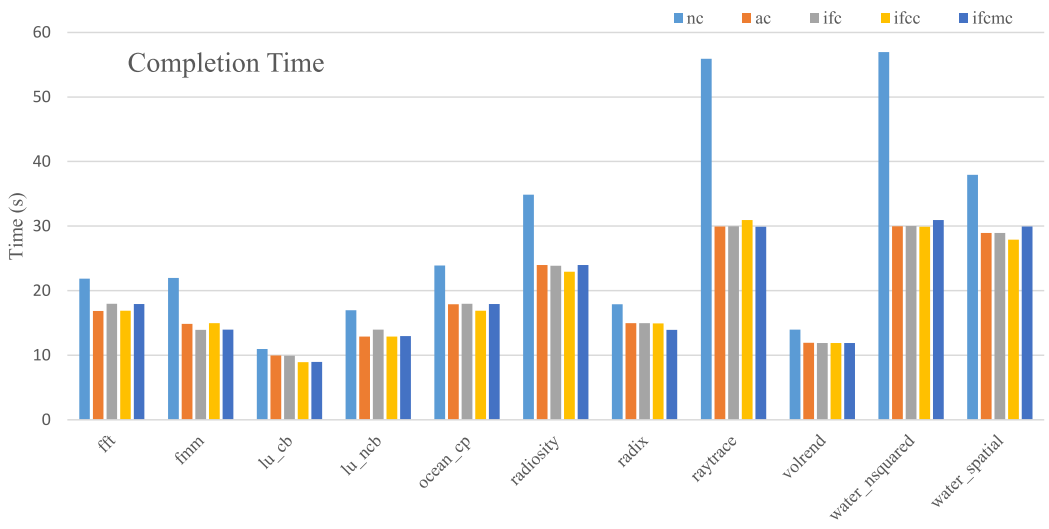


Figure 67: The measurement results for various Splash-2 benchmarks running on the SoC for the completion time.

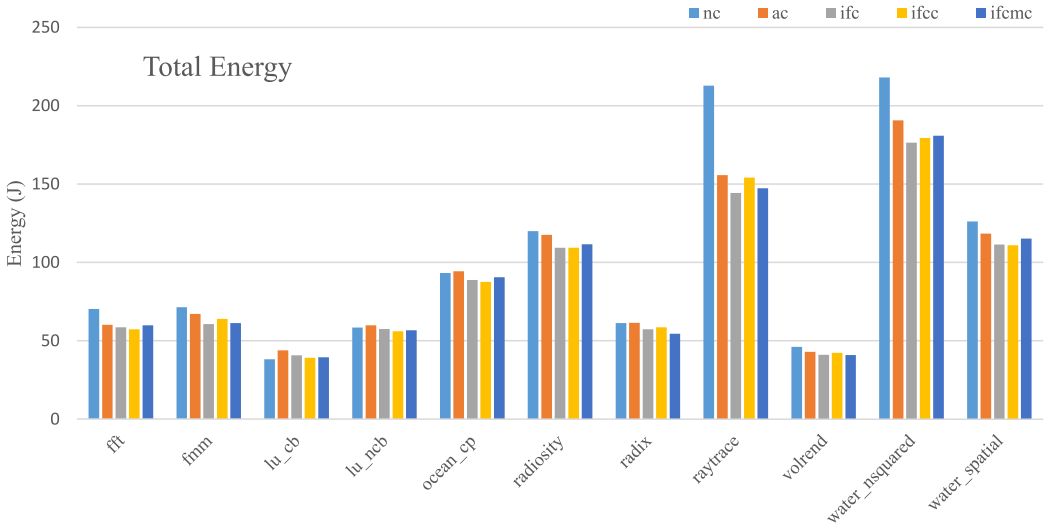


Figure 68: The measurement results for various Splash-2 benchmarks running on the SoC for the total computation energy.

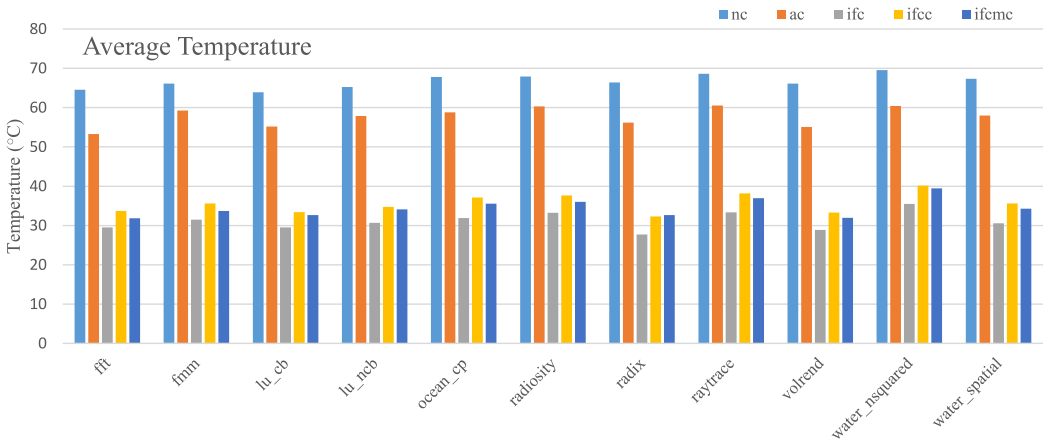


Figure 69: The measurement results for various Splash-2 benchmarks running on the SoC for the average temperature.

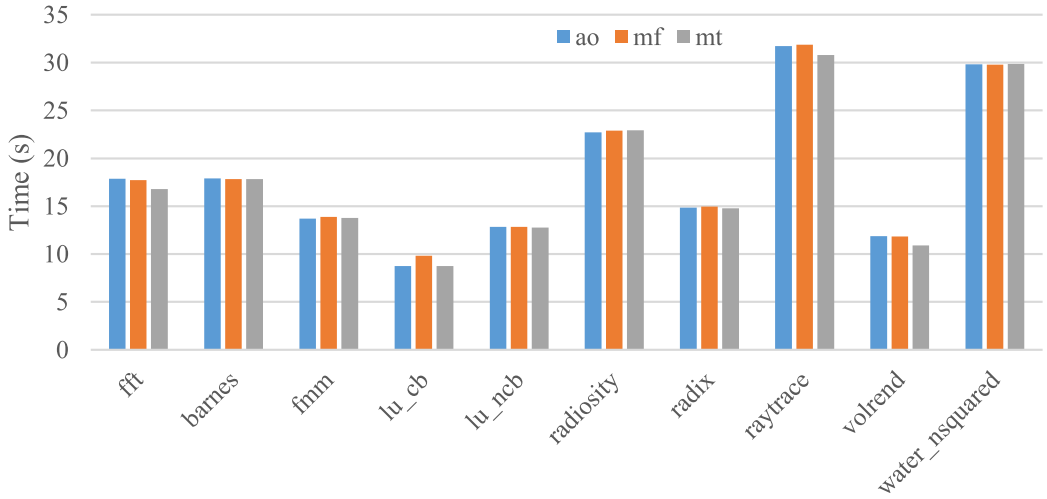


Figure 70: The measurement results for various Splash-2 benchmarks running on the SoC for the completion time. The results highlight the pump enabling policy.

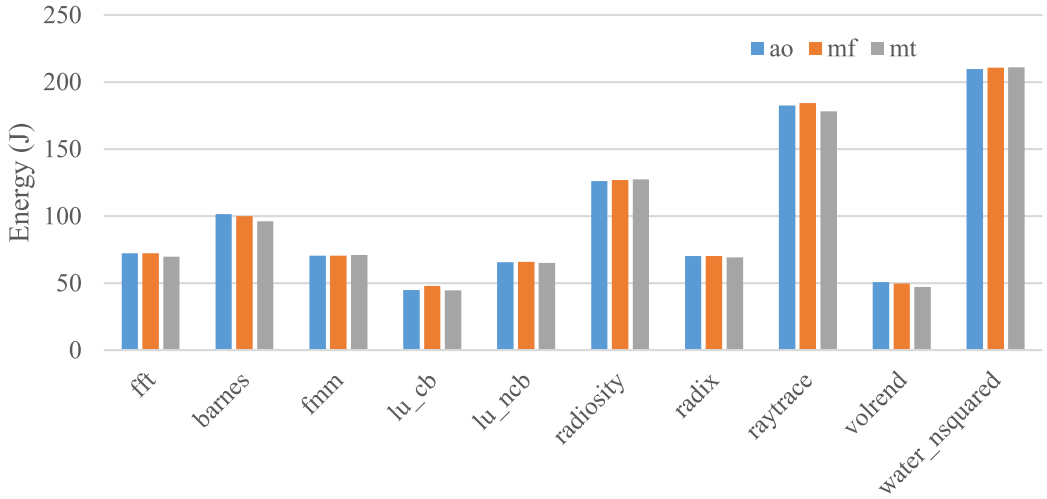


Figure 71: The measurement results for various Splash-2 benchmarks running on the SoC for the total computation energy. The results highlight the pump enabling policy.

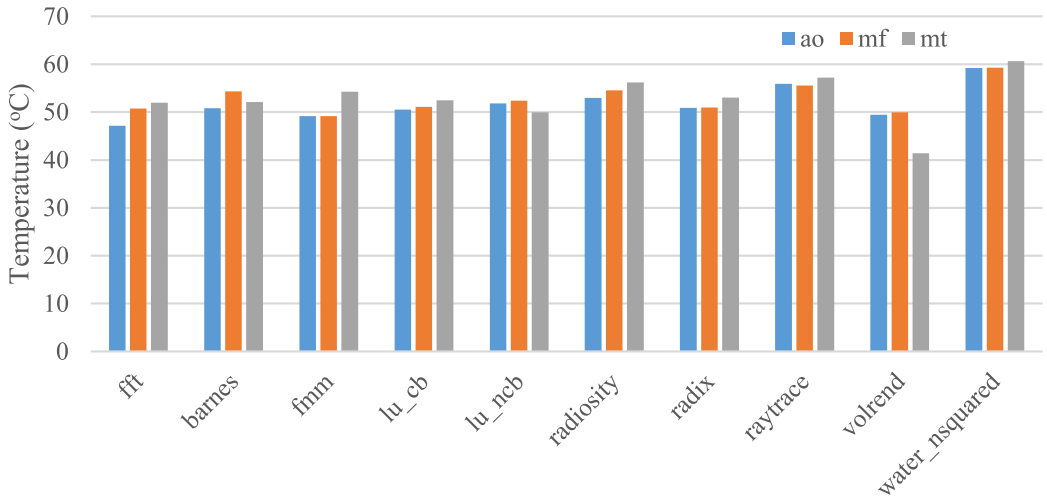


Figure 72: The measurement results for various Splash-2 benchmarks running on the SoC for the average temperature.The results highlight the pump enabling policy.

5.6 Summary

This chapter experimentally demonstrated the application of in-package fluidic cooling for a fully functional commercial SoC. A chip-scale in-package microfluidic cooling technology based on the micro pin fins in a silicon interposer is fabricated and attached to a commercial SoC. The on-board thermal management was demonstrated using a low-power piezoelectric pump controlled by the SoC. The measurements with the benchmark applications showed that compared to the baseline SoC, the in-package cooling achieved 24 % and 16 % lower energy consumptions compared to the baseline (no cooling) and the external passive cooling, respectively. Moreover, the in-package cooling had a reduced assembly footprint and height compared to the external passive and fluidic cooling. Our observation suggested that in mission critical operations when the cores must operate at the maximum load to deliver the required throughput; the in-package cooling solution can successfully complement the electrical techniques (e.g. power gating, voltage-frequency-scaling) to manage temperature and to reduce total system power. The work demonstrated the feasibility of compact chip-scale fluidic cooling structures for SoC without the need for fabricating channels/pin fins directly on the silicon die. We believe, the successful in-package fluidic cooling integration with a commercial SoC will motivate future work on the co-design between innovative cooling structure and advanced SoC power management. The active in-package cooling may bridge efficient fluidic control and workload management to achieve on-line thermal-power co-optimization.

CHAPTER VI

CONCLUSION

6.1 Summary and contribution

In Chapter 2, *we have developed a thermal and supply cross-talks aware performance and robustness analysis methodology for electrical-thermal interaction and applied the methodology to study the 3D processor memory stack*. Our method considers process variation in transistors, thermal field of SRAM tier, power supply variation in the SRAM tier, and the temperature dependency of wire and TSV resistances. The evaluation in the chapter shows that the inter-tier supply and thermal cross-talks may adversely impact the SRAM performance and robustness within a processor-memory stack. Therefore, we conclude that while designing a 3D processormemory stack, the performance and robustness (i.e. parametric failures) of the SRAM array should consider the power dissipation of the processor. Moreover, the thermal cross-talk changes due to the spatiotemporally varying core power and leads to a time varying hot spot radius and intensity. The electrical characteristics of the stacked SRAM will also have corresponding spatiotemporal variations. Our framework identifies these performance and robustness limitations through modeling the thermal field and the power supply conditions. One future direction is to develop better power delivery and cooling technologies to minimize the coupling effects without sacrificing the processor's performance and stability.

In Chapter 3, *we extended the work in Chapter 2 to evaluate EDRAM memory system under electrical-thermal coupling in 2.5 D and 3D environment.* Moreover, the thermal and PDN cross-talks aware performance and robustness analysis methodology for gain cell EDRAM has been extended to model finfet on top of the metal-gate

mosfets. Our method considers the thermal field of memory tier, power supply variation within the tier, and the temperature dependency of wire and TSV resistances. The evaluation shows that the inter-tier supply and thermal cross-talks may adversely impact the EDRAM performance and robustness within a processor-memory stack. The same coupling effect was also in a 2.5D system, but the gradient across the die was more uniform and allows easier design through margin. The horizontal coupling is not of great concern in 2.5D integration, but vertical coupling such as a 3D processormemory stack, the performance and robustness (i.e. parametric failures) of the EDRAM array should consider the power dissipation of the processor.

In Chapter 4, *we developed a methodology to create stimulus and monitoring apparatus on-chip for thermal emulation and thermal field identification – in order to bring advanced thermal testing into the BIST framework.* The framework, field programmable thermal emulator (FPTE) uses on-chip digital heaters and sensors to emulate time-varying power patterns on-chip, generate the corresponding spatiotemporally varying temperature pattern, and characterize the resulting variations in circuit properties, for example, delay. The FPTE may be used as a thermal test-vehicle or as on-line test-structure. The ability to generate any controllable power pattern has been demonstrated to program arbitrary power patterns that the FPTE test-chip emulate the expected power pattern of a target processor/block. The effect of thermal coupling has been demonstrated with the test pattern setup. Having a programmable FPTE test-vehicle allows designers to understand the thermal effect on an architecture and/or workload considering electrical-thermal interactions, but without complete design/fabrication of the functional chip, thereby improve the design turnaround time. As a thermal test-vehicle, FPTE has the advantage over existing thin-film heater based approaches due to its compatibility with standard CMOS process, ability to generate controllable and time-varying power patterns, and directly characterize the effect of temperature patterns on device characteristics. The FPTE

has been used to identify leakage and temperature correlation and transient thermal response for an air cooled system versus a fluidic cooled system. This experiment shows the potential for advanced package evaluation with actual leakage figure and device characteristics. The FPTE has been demonstrated to be programmed on-line and characterize the thermal response of a packaged IC in experiments. This captures the exact thermal characteristics of a specific design-environment interaction considering process variations as well as time-dependent degradation in the thermal properties. Thermal filters were extracted through the FPTE structure for an air filled chip packaging cavity and a fluid filled chip packaging cavity. The recovered filter was used to reconstruct the thermal response of an arbitrarily generated power pattern at an desired location on chip. The correlation of the recovered thermal response and measured thermal response shows close resemblance. This capability allows thermal condition prediction through signature signals, such as the block-level enable signal or clock gating signal, at an arbitrary location on chip. The system allows core level thermal adaptation from signature signals such as power gating or DVFS states.

In Chapter 5, *we apply thermal control structure to an embedded board to improve system power and bandwidth. We experimentally demonstrated the application of in-package fluidic cooling for a fully functional commercial SoC.* A chip-scale in-package microfluidic cooling technology based on the micro pinfins in a silicon interposer is fabricated and attached to a commercial SoC. The on-board thermal management was demonstrated using a low-power piezoelectric pump controlled by the SoC. The measurements with the benchmark applications showed that compared to the baseline SoC, the in-package cooling achieved 24 % and 16 % lower energy consumptions compared to the baseline (no cooling) and the external passive cooling, respectively. Moreover, the in-package cooling had a reduced assembly footprint and height compared to the external passive and fluidic cooling. Our observation suggested that

in mission critical operations when the cores must operate at the maximum load to deliver the required throughput; the in-package cooling solution can successfully complement the electrical techniques (e.g. power gating, voltage-frequency-scaling) to manage temperature and to reduce total system power. The work demonstrated the feasibility of compact chip-scale fluidic cooling structures for SoC without the need for fabricating channels/pinfins directly on the silicon die. We believe, the successful in-package fluidic cooling integration with a commercial SoC will motivate future work on the co-design between innovative cooling structure and advanced SoC power management. The active in-package cooling may bridge efficient fluidic control and workload management to achieve on-line thermal-power co-optimization.

In conclusion, for multi-chip integration, we benefit from exploring electrical-thermal interaction as a whole and co-optimize both components instead of treating temperature and circuit parameters as orthogonal components thats independent to each other. A highly integrated system in an advanced packaging has made the multi-physics interactions increasingly important. The trend of migrating board level circuits onto a system in package (SiP) has been driven by the cost reduction needs in smaller form factor, the power reduction through fewer input-and-output (IO) communications, and a higher communication bandwidth and shorter interconnects between dies within the package. Our journey starts from modeling thermal and electrical interaction for multi-chip systems, through designing test structures for thermal emulation in advanced packages, and ends up at system design for electrical mechanical integration for a functional SoC. The dissertation present a vertically view to electrical-thermal interaction and our share of contribution in each segment.

6.2 Future Work

Our modeling framework in Chapter 2 and Chapter 3 currently does not consider the presence of advanced cooling techniques for 3D ICs – such as liquid cooling and phase

change thermal buffer. Likewise the framework does not consider advanced on-chip voltage regulations. The future work on this direction is to model advanced cooling and power delivery systems and evaluate their impacts on the SRAM performance and robustness. Additionally, a complementary direction will be to explore the opportunity of run-time adaptations considering supply and thermal cross-talks. For example, dynamic thermal managements in 2D multi-core architectures are based on thread migration or core hopping for thermal management. However, in a 3D multi-processor-memory stack, such power migration will also impact the performance of the associated SRAMs and create challenges. Our thermal test structure in Chapter 4 currently does not decode and reconstruct thermal field filter on-chip. The sensor and heater must be programmed from external microcontroller. Ideally the control structure should be build-in and can model and generate/collect complex power pattern on the fly and perform adaptation real-time. This feature will reduce device margin control and build fully adaptive system and squeeze additional performance out of the electrical-thermal coupled environment. Our thermal aware control system in Chapter 5 may be further improved by fully integrated cooling channels as well as piezoelectric pump on the functional die. The highly integrated system may be build with nano-pump thats driven directly inside the fluidic channels from the chip's back end of line or TSVs and allows both air or fluid flow when the local temperature-energy tradeoff is cost-effective. Direct integration also remove additional loss in power regulation and signaling to the pumps. Making active micro-machines directly on top of the integrated SoC has numerous implications on possible advanced integrations. The advantage of such integration may find synergy with various fluidic applications in direct methanol conversion battery, fluidic antenna for wireless signal, lab on chip and bio-inspired nano-vlsi integration.

REFERENCES

[1] *ASTM: Standard Test Method for Thermal Transmission Properties of Thermally Conductive Electrical Insulation Materials*, 2006. Designation D 5470-06.

[2] Al Maashri, A., Sun, G., Dong, X., Narayanan, V., and Xie, Y., "3d gpu architecture using cache stacking: Performance, cost, power and thermal analysis," in *Computer Design, 2009. ICCD 2009. IEEE International Conference on*, pp. 254–259, IEEE, 2009.

[3] Altet, J., Rubio, A., Schaub, E., Dilhaire, S., and Claeys, W., "Thermal coupling in integrated circuits: application to thermal testing," *Solid-State Circuits, IEEE Journal of*, vol. 36, no. 1, pp. 81–91, 2001.

[4] Andric, A. and Wigton, D. L., "Split power supply subsystem with isolated voltage supplies to satisfy a predetermined power limit," Nov. 13 2007. US Patent 7,294,976.

[5] Bar-Cohen, A., "Thermal management of air-and liquid-cooled multichip modules," *Components, Hybrids, and Manufacturing Technology, IEEE Transactions on*, vol. 10, no. 2, pp. 159–175, 1987.

[6] Benedek, Z., Courtois, B., Farkas, G., KOLLar, E., Mir, S., Poppe, A., Rencz, M., Székely, V., and Torki, K., "A scalable multi-functional thermal test chip family: Design and evaluation," *Journal of Electronic Packaging*, vol. 123, no. 4, pp. 323–330, 2001.

[7] Borkar, S., "Thousand core chips: a technology perspective," in *Proceedings of the 44th annual Design Automation Conference*, pp. 746–749, ACM, 2007.

[8] Brooks, D. and Martonosi, M., "Dynamic thermal management for high-performance microprocessors," in *High-Performance Computer Architecture, 2001. HPCA. The Seventh International Symposium on*, pp. 171–182, IEEE, 2001.

[9] Chatterjee, S., Cho, M., Rao, R., and Mukhopadhyay, S., "Impact of die-to-die thermal coupling on the electrical characteristics of 3d stacked sram cache," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2012 28th Annual IEEE*, pp. 14–19, IEEE, 2012.

[10] Chen, C.-C. and Milor, L., "Microprocessor aging analysis and reliability modeling due to back-end wearout mechanisms," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2014.

[11] Cheng, Y.-K., *Electrothermal analysis of VLSI systems*. Springer Science & Business Media, 2000.

[12] Cho, M., Khellah, M., Chae, K., Ahmed, K., Tschanz, J., and Mukhopadhyay, S., "Characterization of inverse temperature dependence in logic circuits," in *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*.

[13] Cho, M., Song, W., Yalamanchili, S., and Mukhopadhyay, S., "Thermal system identification (tsi): A methodology for post-silicon characterization and prediction of the transient thermal field in multicore chips," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2012 28th Annual IEEE*, pp. 118–124, IEEE, 2012.

[14] Chun, K. C., Jain, P., Kim, T.-H., and Kim, C. H., "A 667 mhz logic-compatible embedded dram featuring an asymmetric 2t gain cell for high speed on-die caches," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 2, pp. 547–559, 2012.

[15] Chun, K. C., Jain, P., Lee, J. H., and Kim, C. H., "A 3t gain cell embedded dram utilizing preferential boosting for high density and low power on-die caches," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 6, pp. 1495–1505, 2011.

[16] Cochran, R. and Reda, S., "Spectral techniques for high-resolution thermal characterization with limited sensor data," in *Proceedings of the 46th Annual Design Automation Conference*, pp. 478–483, ACM, 2009.

[17] Colinge, J.-P., Floyd, L., Quinn, A. J., Redmond, G., Alderman, J. C., Xiong, W., Cleavelin, C. R., Schulz, T., Schruefer, K., Knoblinger, G., and others, "Temperature effects on trigate soi mosfets," *Electron Device Letters, IEEE*, vol. 27, no. 3, pp. 172–174, 2006.

[18] Diodato, P. W., "Embedded dram: more than just a memory," *Communications Magazine, IEEE*, vol. 38, no. 7, pp. 118–126, 2000.

[19] Dorsey, P., "Xilinx stacked silicon interconnect technology delivers breakthrough fpga capacity, bandwidth, and power efficiency," *Xilinx White Paper: Virtex-7 FPGAs*, pp. 1–10, 2010.

[20] E&CE, S. and USIT, G., "Static noise margin analysis of sram cell for high speed application," *IJCSI*, p. 175, 2010.

[21] Ellsworth, M. J., Campbell, L., Simons, R., and Iyengar, R., "The evolution of water cooling for ibm large server systems: Back to the future," in *Thermal and Thermomechanical Phenomena in Electronic Systems, 2008. ITHERM 2008. 11th Intersociety Conference on*, pp. 266–274, IEEE, 2008.

[22] George, V., Jahagirdar, S., Tong, C., Smits, K., Damaraju, S., Siers, S., Naydenov, V., Khondker, T., Sarkar, S., and Singh, P., "Penryn: 45-nm next generation intel® core 2 processor," in *Solid-State Circuits Conference, 2007. ASSCC'07. IEEE Asian*, pp. 14–17, IEEE, 2007.

[23] Grossar, E., Stucchi, M., Maex, K., and Dehaene, W., "Read stability and write-ability analysis of sram cells for nanometer technologies," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 11, pp. 2577–2588, 2006.

[24] Gubbi, J., Buyya, R., Marusic, S., and Palaniswami, M., "Internet of things (iot): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[25] Gupta, M. S., Oatley, J. L., Joseph, R., Wei, G.-Y., and Brooks, D. M., "Understanding voltage variations in chip multiprocessors using a distributed power-delivery network," in *Design, Automation & Test in Europe Conference & Exhibition, 2007. DATE'07*, pp. 1–6, IEEE, 2007.

[26] Gupta, S. S., Siva, V. M., Krishnan, S., Sreeprasad, T., Singh, P. K., Pradeep, T., and Das, S. K., "Thermal conductivity enhancement of nanofluids containing graphene nanosheets," *Journal of Applied Physics*, vol. 110, no. 8, p. 084302, 2011.

[27] Gurrum, S. P., Edwards, D. R., Marchand-Golder, T., Akiyama, J., Yokoya, S., Drouard, J.-F., and Dahan, F., "Generic thermal analysis for phone and tablet systems," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, pp. 1488–1492, IEEE, 2012.

[28] Hamburgen, W. R. and Fitch, J. S., "Packaging a 150-w bipolar ecl microprocessor," *Components, Hybrids, and Manufacturing Technology, IEEE Transactions on*, vol. 16, no. 1, pp. 28–38, 1993.

[29] Hamzaoglu, F., Arslan, U., Bisnik, N., Ghosh, S., Lal, M. B., Lindert, N., Meterelliyoz, M., Osborne, R. B., Park, J., Tomishima, S., and others, "13.1 a 1gb 2ghz embedded dram in 22nm tri-gate cmos technology," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, pp. 230–231, IEEE, 2014.

[30] Healy, M. B. and Lim, S. K., "A novel tsv topology for many-tier 3d power-delivery networks," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011*, pp. 1–4, IEEE, 2011.

[31] Hua, H., Mineo, C., Schoenfliess, K., Sule, A., Melamed, S., Jenkal, R., and Davis, W. R., "Exploring compromises among timing, power and temperature in three-dimensional integrated circuits," in *Proceedings of the 43rd annual Design Automation Conference*, pp. 997–1002, ACM, 2006.

[32] Huang, G., Bakir, M., Naeemi, A., Chen, H., and Meindl, J. D., "Power delivery for 3d chip stacks: Physical modeling and design implication," in *Electrical Performance of Electronic Packaging, 2007 IEEE*, pp. 205–208, IEEE, 2007.

[33] Jacob, P., Erdogan, O., Zia, A., Belemjian, P. M., Kraft, R. P., and McDonald, J. F., "Predicting the performance of a 3d processor-memory chip stack," *Design & Test of Computers, IEEE*, vol. 22, no. 6, pp. 540–547, 2005.

[34] Janicki, M., Collet, J. H., Louri, A., and Napieralski, A., "Hot spots and core-to-core thermal coupling in future multi-core architectures," in *Semiconductor Thermal Measurement and Management Symposium, 2010. SEMI-THERM 2010. 26th Annual IEEE*, pp. 205–210, IEEE, 2010.

[35] Johnson, L., "Sony answers 4k overheating concerns,...," 2014.

[36] Katti, G., Stucchi, M., De Meyer, K., and Dehaene, W., "Electrical modeling and characterization of through silicon via for three-dimensional ics," *Electron Devices, IEEE Transactions on*, vol. 57, no. 1, pp. 256–262, 2010.

[37] Khan, N. H., Alam, S. M., and Hassoun, S., "System-level comparison of power delivery design for 2d and 3d ics," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, pp. 1–7, IEEE, 2009.

[38] Kim, S. Y. and Webb, R. L., "Analysis of convective thermal resistance in ducted fan-heat sinks," *Components and Packaging Technologies, IEEE Transactions on*, vol. 29, no. 3, pp. 439–448, 2006.

[39] Kim, W., Gupta, M. S., Wei, G.-Y., and Brooks, D., "System level analysis of fast, per-core dvfs using on-chip switching regulators," in *High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on*, pp. 123–134, IEEE, 2008.

[40] Knickerbocker, J., Andry, P., Colgan, E., Dang, B., Dickson, T., Gu, X., Haymes, C., Jahnes, C., Liu, Y., Maria, J., and others, "2.5 d and 3d technology challenges and test vehicle demonstrations," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, pp. 1068–1076, IEEE, 2012.

[41] Kuhn, K. J., "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale cmos," in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 471–474, IEEE, 2007.

[42] Kuhn, K. J., Giles, M. D., Becher, D., Kolar, P., Kornfeld, A., Kotlyar, R., Ma, S. T., Maheshwari, A., and Mudanai, S., "Process technology variation," *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2197–2208, 2011.

[43] Kumagai, K., Yang, C., Izumino, H., Narita, N., Shinjo, K., Iwashita, S.-i., Nakaoka, Y., Kawamura, T., Komabashiri, H., Minato, T., and others, "System-in-silicon architecture and its application to h. 264/avc motion estimation for 1080hdtv," in *Solid-state circuits conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, pp. 1706–1715, IEEE, 2006.

[44] Kung, J., Yueh, W., Yalamanchili, S., and Mukhopadhyay, S., "Post-silicon estimation of spatiotemporal temperature variations using mimo thermal filters," vol. 5, pp. 650–660, May 2015.

[45] Kurabayashi, K. and Goodson, K. E., "Precision measurement and mapping of die-attach thermal resistance," *Components, Packaging, and Manufacturing Technology, Part A, IEEE Transactions on*, vol. 21, no. 3, pp. 506–514, 1998.

[46] Kurd, N., Chowdhury, M., Burton, E., Thomas, T. P., Mozak, C., Boswell, B., Mosalikanti, P., Neidengard, M., Deval, A., Khanna, A., and others, "Haswell: A family of ia 22 nm processors," *Solid-State Circuits, IEEE Journal of*, vol. 50, no. 1, pp. 49–58, 2015.

[47] Lau, J. H. and others, "Tsv interposers: The most cost-effective integrator for 3d ic integration," *Chip Scale Rev*, vol. 15, no. 5, pp. 23–27, 2011.

[48] Lee, D. U., Kim, K. W., Kim, K. W., Kim, H., Kim, J. Y., Park, Y. J., Kim, J. H., Kim, D. S., Park, H. B., Shin, J. W., and others, "25.2 a 1.2 v 8gb 8-channel 128gb/s high-bandwidth memory (hbm) stacked dram with effective microbump i/o test methods using 29nm process and tsv," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, pp. 432–433, IEEE, 2014.

[49] Lee, J., Gerlach, D. W., and Joshi, Y. K., "Parametric thermal modeling of heat transfer in handheld electronic devices," in *Thermal and Thermomechanical Phenomena in Electronic Systems, 2008. ITHERM 2008. 11th Intersociety Conference on*, pp. 604–609, IEEE, 2008.

[50] Loi, G. L., Agrawal, B., Srivastava, N., Lin, S.-C., Sherwood, T., and Banerjee, K., "A thermally-aware performance analysis of vertically integrated (3-d) processor-memory hierarchy," in *Proceedings of the 43rd annual Design Automation Conference*, pp. 991–996, ACM, 2006.

[51] Luk, W. K., Cai, J., Dennard, R. H., Immediato, M. J., and Kosonocky, S. V., "A 3-transistor dram cell with gated diode for enhanced speed and retention time," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, pp. 184–185, IEEE, 2006.

[52] Luk, W. K. and Dennard, R. H., "A novel dynamic memory cell with internal voltage gain," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 4, pp. 884–894, 2005.

[53] Mukhopadhyay, S., Kim, K., Mahmoodi, H., and Roy, K., "Design of a process variation tolerant self-repairing sram for yield enhancement in nanoscaled cmos," *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 6, pp. 1370–1382, 2007.

[54] Mukhopadhyay, S., Mahmoodi, H., and Roy, K., "Modeling of failure probability and statistical design of sram array for yield enhancement in nanoscaled cmos," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 24, no. 12, pp. 1859–1880, 2005.

[55] Mukhopadhyay, S., Rao, R. M., Kim, J.-J., and Chuang, C.-T., "Sram write-ability improvement with transient negative bit-line voltage," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 1, pp. 24–32, 2011.

[56] Mukhopadhyay, S., Raychowdhury, A., and Roy, K., "Accurate estimation of total leakage in nanometer-scale bulk cmos circuits based on device geometry and doping profile," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 24, no. 3, pp. 363–381, 2005.

[57] Nagy, G., Horvath, P., and Poppe, A., "Practical aspects of thermal transient testing in live digital circuits," in *Thermal Investigations of ICs and Systems (THERMINIC), 2013 19th International Workshop on*, pp. 87–91, IEEE, 2013.

[58] Norem, J., "Amd unleashes the dual-gpu radeon r9 295x2," 2014.

[59] Ozdemir, S., Pan, Y., Das, A., Memik, G., Loh, G., and Choudhary, A., "Quantifying and coping with parametric variations in 3d-stacked microarchitectures," in *Proceedings of the 47th Design Automation Conference*, pp. 144–149, ACM, 2010.

[60] Pak, J. S., Kim, J., Cho, J., Kim, K., Song, T., Ahn, S., Lee, J., Lee, H., Park, K., and Kim, J., "Pdn impedance modeling and analysis of 3d tsv ic by using proposed p/g tsv array model based on separated p/g tsv and chip-pdn models," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 1, no. 2, pp. 208–219, 2011.

[61] Park, Y. S., Blaauw, D., Sylvester, D., and Zhang, Z., "A 1.6-mm 2 38-mw 1.5-gb/s ldpc decoder enabled by refresh-free embedded dram," in *VLSI Circuits (VLSIC), 2012 Symposium on*, pp. 114–115, IEEE, 2012.

[62] Peles, Y., Koşar, A., Mishra, C., Kuo, C.-J., and Schneider, B., "Forced convective heat transfer across a pin fin micro heat sink," *International Journal of Heat and Mass Transfer*, vol. 48, no. 17, pp. 3615–3627, 2005.

[63] Poppe, A. and Székely, V., "Dynamic temperature measurements: tools providing a look into package and mount structures," *Electronics Cooling*, vol. 8, pp. 10–19, 2002.

[64] Puttaswamy, K. and Loh, G. H., "Implementing caches in a 3d technology for high performance processors," in *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, pp. 525–532, IEEE, 2005.

[65] Reda, S., "Thermal and power characterization of real computing devices," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 1, no. 2, pp. 76–87, 2011.

[66] Roy, K., Mukhopadhyay, S., and Mahmoodi-Meimand, H., "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003.

[67] Sekar, D., King, C., Dang, B., Spencer, T., Thacker, H., Joseph, P., Bakir, M., and Meindl, J., "A 3d-ic technology with integrated microchannel cooling," in *Interconnect Technology Conference, 2008. IITC 2008. International*, pp. 13–15, IEEE, 2008.

[68] Semiconductor, T., "Tezzaron unveils 3d sram," 2005.

[69] Serafy, C., Srivastava, A., and Yeung, D., "Continued frequency scaling in 3d ics through micro-fluidic cooling," in *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014 IEEE Intersociety Conference on*, pp. 79–85, IEEE, 2014.

[70] Shao, L., Raghavan, A., Emurian, L., Papaefthymiou, M. C., Wenisch, T. F., Martin, M. M., and Pipe, K. P., "On-chip phase change heat sinks designed for computational sprinting," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2014 30th Annual*, pp. 29–34, IEEE, 2014.

[71] Shayan, A., Hu, X., Zhang, W., Cheng, C.-K., Chen, X., Popovich, M., and others, "3d stacked power distribution considering substrate coupling," in *Computer Design, 2009. ICCD 2009. IEEE International Conference on*, pp. 225–230, IEEE, 2009.

[72] Siegal, B. and Galloway, J., "Thermal test chip design and performance considerations," in *Semiconductor Thermal Measurement and Management Symposium, 2008. Semi-Therm 2008. Twenty-fourth Annual IEEE*, pp. 59–62, IEEE, 2008.

[73] Sinha, S., Yeric, G., Chandra, V., Cline, B., and Cao, Y., "Exploring sub-20nm finfet design with predictive technology models," in *Proceedings of the 49th Annual Design Automation Conference*, pp. 283–288, ACM, 2012.

[74] Smith, S. E. and Campbell, R. C., "Flash diffusivity method: A survey of capabilities," *ElectronicsCoolings, May*, 2002.

[75] Song, W., Mukhopadhyay, S., and Yalamanchili, S., "Architectural reliability: Lifetime reliability characterization and management of many-core processors," vol. PP, no. 99, pp. 1–1, 2014.

[76] Sridhar, A., Vincenzi, A., Ruggiero, M., Brunschwiler, T., and Atienza, D., "3d-ice: Fast compact transient thermal modeling for 3d ics with inter-tier liquid cooling," in *Proceedings of the International Conference on Computer-Aided Design*, pp. 463–470, IEEE Press, 2010.

[77] Sun, G., Wu, X., and Xie, Y., "Exploration of 3d stacked l2 cache design for high performance and efficient thermal control," in *Proceedings of the 2009 ACM/IEEE international symposium on Low power electronics and design*, pp. 295–298, ACM, 2009.

[78] Tarter, T. S., "Programming thermal test chip arrays," May 6 2003. US Patent 6,559,667.

[79] Teman, A., Meinerzhagen, P. A., Burg, A. P., and Fish, A., "Review and classification of gain cell edram implementations," in *IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, no. EPFL-REVIEW-181635, 2012.

[80] Topol, A. W., La Tulipe, D., Shi, L., Frank, D. J., Bernstein, K., Steen, S. E., Kumar, A., Singco, G. U., Young, A. M., Guarini, K. W., and others, "Three-dimensional integrated circuits," *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 491–506, 2006.

[81] Trivedi, A. R., Yueh, W., and Mukhopadhyay, S., "Impact of through-silicon-via capacitance on high frequency supply noise in 3d-stacks," in *Electrical Performance of Electronic Packaging and Systems (EPEPS), 2011 IEEE 20th Conference on*, pp. 105–108, IEEE, 2011.

[82] Tsioutsios, I., Pavlidis, V. F., and De Micheli, G., "Physical design tradeoffs in power distribution networks for 3-d ics," in *Electronics, Circuits, and Systems (ICECS), 2010 17th IEEE International Conference on*, pp. 430–433, IEEE, 2010.

[83] Tuckerman, D. B. and Pease, R., "High-performance heat sinking for vlsi," *Electron Device Letters, IEEE*, vol. 2, no. 5, pp. 126–129, 1981.

[84] Uchida, H., Shioga, T., Aoki, S., Ogata, S., and Nagaoka, H., "Loop heat pipe," Aug. 22 2012. US Patent App. 13/591,397.

[85] Wagner, G. R. and Maltz, W., "On the thermal management challenges in next generation handheld devices," in *ASME 2013 International Technical Conference and Exhibition on Packaging and Integration of Electronic and*

Photonic Microsystems, pp. V002T08A046–V002T08A046, American Society of Mechanical Engineers, 2013.

[86] Wan, Z., Yueh, W., Joshi, Y., and Mukhopadhyay, S., "Enhancement in cmos chip performance through microfluidic cooling," in *Thermal Investigations of ICs and Systems (THERMINIC), 2014 20th International Workshop on*, pp. 1–5, IEEE, 2014.

[87] Wang, J., Nalam, S., and Calhoun, B. H., "Analyzing static and dynamic write margin for nanometer srams," in *Low Power Electronics and Design (ISLPED), 2008 ACM/IEEE International Symposium on*, pp. 129–134, IEEE, 2008.

[88] Ward, B. C., Herman, J. L., Kenna, C. J., and Anderson, J. H., "Outstanding paper award: Making shared caches more predictable on multicore platforms," in *Real-Time Systems (ECRTS), 2013 25th Euromicro Conference on*, pp. 157–167, IEEE, 2013.

[89] Waterland, A., "Stress," 2014.

[90] Weber, P., Zagrabski, M., Wojciechowski, B., Berezowski, K. S., Nikodem, M., and Kepa, K., "Toolset for measuring thermal behavior of fpga devices," in *Thermal Investigations of ICs and Systems (THERMINIC), 2013 19th International Workshop on*, pp. 48–53, IEEE, 2013.

[91] Woo, S. C., Ohara, M., Torrie, E., Singh, J. P., and Gupta, A., "The splash-2 programs: Characterization and methodological considerations," in *ACM SIGARCH Computer Architecture News*, vol. 23, pp. 24–36, ACM, 1995.

[92] Xiao, H., Wan, Z., Yalamanchili, S., and Joshi, Y., "Leakage power characterization and minimization in 3d stacked multi-core chips with microfluidic cooling," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2014 30th Annual*, pp. 207–212, IEEE, 2014.

[93] Yueh, W., Chatterjee, S., Trivedi, A. R., and Mukhopadhyay, S., "On the parametric failures of sram in a 3d-die stack considering tier-to-tier supply cross-talk.," in *VTS*, pp. 264–269, 2012.

[94] Yun, W., Kang, K., and Kyung, C.-M., "Thermal-aware energy minimization of 3d-stacked l3 cache with error rate limitation," in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pp. 1672–1675, IEEE, 2011.

[95] Zhan, Y. and Sapatnekar, S. S., "High-efficiency green function-based thermal simulation algorithms," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 9, pp. 1661–1675, 2007.

[96] Zhang, H. and Balog, R. S., "Loss analysis during dead time and thermal study of gallium nitride devices," in *Applied Power Electronics Conference and Exposition (APEC), 2015 IEEE*, pp. 737–744, IEEE, 2015.

[97] Zhang, Y. and Bakir, M. S., "Independent interlayer microfluidic cooling for heterogeneous 3d ic applications," *Electronics Letters*, vol. 49, no. 6, pp. 404–406, 2013.

[98] Zhang, Y., Dembla, A., Joshi, Y., and Bakir, M. S., "3d stacked microfluidic cooling for high-performance 3d ics," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, pp. 1644–1650, IEEE, 2012.

[99] Zhao, J., Sun, G., Loh, G. H., and Xie, Y., "Optimizing gpu energy efficiency with 3d die-stacking graphics memory and reconfigurable memory interface," *ACM Transactions on Architecture and Code Optimization (TACO)*, vol. 10, no. 4, p. 24, 2013.

[100] Zhao, W. and Cao, Y., "New generation of predictive technology model for sub-45 nm early design exploration," *Electron Devices, IEEE Transactions on*, vol. 53, no. 11, pp. 2816–2823, 2006.

[101] Zhou, P., Sridharan, K., and Sapatnekar, S. S., "Congestion-aware power grid optimization for 3d circuits using mim and cmos decoupling capacitors," in *Proceedings of the 2009 Asia and South Pacific Design Automation Conference*, pp. 179–184, IEEE Press, 2009.

VITA

Wen Yueh (S' 08) received his B.S. and M.S. degrees in Electrical and Computer Engineering from Rutgers University, New Jersey, in 2009. He received his Ph.D. degree in Georgia Institute of Technology, Atlanta, in 2015. His research interest includes system level multi-physics simulator, self-adaptive circuit design for many-core processor thermal management, and energy-aware low power memory architecture.